Fundamental Frequency and Direction-of-Arrival Estimation for Multichannel Speech Enhancement

Karimian-Azari, Sam

DOI (link to publication from Publisher):
10.5278/vbn.phd.engsci.00126

Publication date:
2016

Document Version
Publisher's PDF, also known as Version of record

FUNDAMENTAL FREQUENCY AND DIRECTION-OF-ARRIVAL ESTIMATION FOR MULTICHANNEL SPEECH ENHANCEMENT

BY
SAM KARIMIAN-AZARI

DISSERTATION SUBMITTED 2016

AALBORG UNIVERSITY
DENMARK

Fundamental Frequency and Direction-of-Arrival Estimation for Multichannel Speech Enhancement

PhD Dissertation

Sam Karimian-Azari

Department of Architecture, Design and Media Technology
Aalborg University
Rendsbuggade 14
DK–9000 Aalborg
Denmark

Thesis submitted: June 2016

PhD Supervisor: Prof. Mads Græsbøll Christensen
Aalborg University

PhD Co-supervisor: Postdoc. Jesper Rindom Jensen
Aalborg University

PhD Committee: Assoc. Prof. Kamal Nasrollahi (chairman)
Aalborg University

Prof. Augusto Sarti
Polytechnic University of Milan

Assoc. Prof. Roland Badeau
Telecom ParisTech

PhD Series: Faculty of Engineering and Science,
Aalborg Univer-sity

ISSN (online): 2246-1248
ISBN (online): 978-87-7112-750-8

Published by:
Aalborg University Press
Skjernvej 4A, 2nd floor
DK–9220 Aalborg Ø
Phone: +45 99407140
aauf@forlag.aau.dk
forlag.aau.dk

Printed in Denmark by Rosendahls, 2016

Abstract

Audio systems receive the speech signals of interest usually in the presence of noise. The noise has profound impacts on the quality and intelligibility of the speech signals, and it is therefore clear that the noisy signals must be cleaned up before being played back, stored, or analyzed. We can estimate the speech signal of interest from the noisy signals using a priori knowledge about it. A human speech signal is broadband and consists of both voiced and unvoiced parts. The voiced part is quasi-periodic with a time-varying fundamental frequency (or pitch as it is commonly referred to). We consider the periodic signals basically as the sum of harmonics. Therefore, we can pass the noisy signals through bandpass filters centered at the frequencies of the harmonics to enhance the signal. In addition, although the frequencies of the harmonics are the same across the channels of a microphone array, the multichannel periodic signals may have different phases due to the time-differences-of-arrivals (TDOAs) which are related to the direction-of-arrival (DOA) of the impinging sound waves. Hence, the outputs of the array can be steered to the direction of the signal of interest in order to align their time differences which eventually may further reduce the effects of noise.

This thesis introduces a number of principles and methods to estimate periodic signals in noisy environments with application to multichannel speech enhancement. We propose model-based signal enhancement concerning the model of periodic signals. Therefore, the parameters of the model must be estimated in advance. The signal of interest is often contaminated by different types of noise that may render many estimation methods suboptimal due to an incorrect white Gaussian noise assumption. We therefore propose robust estimators against the noise and focus on statistical-based and filtering-based methods by imposing distortionless constraints with explicit relations between the parameters of the harmonics. The estimated fundamental frequencies are expected to be continuous over time. Therefore, we concern the time-varying fundamental frequency in the statistical methods in order to lessen the estimation error. We also propose a maximum likelihood DOA estimator concerning the noise statistics and the linear relationship between the TDOAs of the harmonics. The estimators have benefits compared

to the state-of-the-art statistical-based methods in colored noise. Evaluations of the estimators comparing with the minimum variance of the deterministic parameters and the other methods confirm that the proposed estimators are statistically efficient in colored noise and computationally simple. Finally, we propose model-based beamformers in multichannel speech signal enhancement by exploiting the estimated fundamental frequency and DOA of the signal of interest. This general framework is tailored to a number of beamformers concerning the spectral and spatial information of the periodic signals which are quasi-stationary in short intervals. Objective measures of speech quality and ineligibility confirm the advantage of the harmonic model-based beamformers over the traditional beamformers, which are non-parametric, and reveal the importance of an accurate estimate of the parameters of the model.

Resumé

Når et lydsystem modtager et talesignal, vil dette oftest indeholde støj. Støjen har en markant indvirkning på talesignalets kvalitet og forståelighed, hvorfor det er nødvendigt at fjerne støjsignalerne, før en optagelse afspilles, lagres eller analyseres. Forudgående viden kan anvendes til at adskille de ønskede talesignaler fra støjsignalerne. Talesignaler er bredbåndet er bredbåndet og består af både stemte og ustemte lyde. Det stemte talesignal er kvasi-periodisk og har en grundfrekvens, som varierer over tid (og som normalt omtales som toneleje). Periodiske signaler kan ses som en sum af harmoniske komponenter. Derfor kan vi lade støjsignalerne passere igennem båndpasfiltre, der forstærker signalet ved at fokusere på de harmoniske komponenters frekvenser. Selvom de harmoniske komponenters frekvenser er identiske på tværs af et mikrofonarrays kanaler, kan multikanals periodiske signaler udgøres af forskellige faser på grund af time-differences-of-arrivals (TDOA), som hænger sammen med lydbølgernes direction-of-arrival (DOA). Arrayets output kan derfor ledes i retning af det ønskede signal, og dermed bliver det muligt at afstemme de tidsforskelle, der kan reducere følgevirkningerne af støj yderligere.

Denne afhandling introducerer en række principper og metoder til estimeringen af periodiske signaler i støjende miljøer med anvendelse af multikanals støjreduktion af talesignaler. Vi foreslår, at modelbaseret støjreduktion til periodiske signaler. Derfor bør modellens parametre estimeres på forhånd. Det ønskede signal forurenes ofte af forskellige former for støj, hvilket kan bevirke, at mange estimeringsmetoder fejler pga en fejlagtig antagelse om hvis gaussisk støj. Vi foreslår derfor anvendelsen af estimatorer robuste mod støj med fokus på statistik- og filtreringsbaserede metoder, der lægger forvrængningsfri sidebetingelser med direkte forbindelse imellem parametrene for de harmoniske komponenter. Det forventes, at den estimerede grundfrekvens er kontinuert over tid. For at mindske risikoen for estimeringsfejl, betragter vi derfor den tidsvarierende grundfrekvens i de statistiske metoder. Vi foreslår desuden en maximum likelihood DOA estimator baseret på støjstatistikker og den lineære forbindelse mellem TDOA'erne af de harmoniske komponenter. Sammenlignet med de seneste statistisk baserede

metoder til brug ved farvet støj, har disse metoder visse fordele. Evalueringer af estimatorerne sammenholdt med den minimale varians for de deterministiske parametre og andre metoder bekræfter, at de foreslåede estimatorer statistisk set er effektive i miljøer med farvet støj og simple set ud fra et beregningsteknisk perspektiv. Endelig foreslår vi brugen af modelbaserede beamformere i multikanals støjreduktion af talesignaler til optimal udnyttelse af det ønskede signals estimerede grundfrekvens og DOA. Indenfor dette framework udvikles en række beamformere, der rettes imod den spektrale og rumlige information indeholdt i de periodiske signaler, som er kvasi-stationære i korte intervaller. Objektive mål for talekvalitet og -forståelighed bekræfter, at harmoniske modelbaserede beamformere er fordelagtige sammenlignet med traditionelle beamformere, da sidstnævnte ikke er parametriske og demonstrerer fordelen ved at kunne finde præcise parameterestimater.

Contents

Thesis Details

Thesis Title:	Fundamental Frequency and Direction-of-Arrival Estimation for Multichannel Speech Enhancement
PhD Student:	Sam Karimian-Azari
Supervisor:	Prof. Mads Græsbøll Christensen, Aalborg University
Co-supervisor:	Postdoc. Jesper Rindom Jensen, Aalborg University

The main body of this thesis consists of the following (peer-reviewed) papers.

[A] S. Karimian-Azari, J. R. Jensen, J. Benesty, M. G. Christensen, "A class of parametric broadband beamformers based on the fundamental frequency," *J. Acoust. Soc. Am.*, 2016 (submitted).

[B] S. Karimian-Azari, J. R. Jensen, M. G. Christensen, "Computationally efficient and noise robust DOA and pitch estimation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 24, no. 9, pp. 1609–1621, 2016.

[C] S. Karimian-Azari, A. Jakobsson, J. R. Jensen, M. G. Christensen, "Multi-pitch estimation and tracking using Bayesian inference in block sparsity," *Proc. European Signal Processing Conf.*, pp. 16–20, 2015.

[D] S. Karimian-Azari, N. Mohammadiha, J. R. Jensen, M. G. Christensen, "Pitch estimation and tracking with harmonic emphasis on the acoustic spectrum," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 4330–4334, 2015.

[E] S. Karimian-Azari, J. R. Jensen, M. G. Christensen, "Fast joint DOA and pitch estimation using a broadband MVDR beamformer," *Proc. European Signal Processing Conf.*, pp. 1–5, 2013.

[F] S. Karimian-Azari, J. R. Jensen, M. G. Christensen, "Fundamental frequency and model order estimation using spatial filtering," *Proc. IEEE Int. Confer. Acoust., Speech, Signal Process.*, pp. 5964–5968, 2014.

[G] S. Karimian-Azari, J. Benesty, J. R. Jensen, M. G. Christensen, "A broadband beamformer using controllable constraints and minimum variance," *Proc. European Signal Processing Conf.*, pp. 666–670, 2014.

In addition to the main papers, the following publications have also been made.

[1] S. Karimian-Azari, J. R. Jensen, M. G. Christensen, "Robust pitch estimation using an optimal filter on frequency estimates," *Proc. European Signal Processing Conf.*, pp. 1557–1561, 2014.

[2] S. Karimian-Azari, J. R. Jensen, M. G. Christensen, "Robust DOA estimation of harmonic signals using constrained filters on phase estimates," *Proc. European Signal Processing Conf.*, pp. 1930–1934, 2014.

This thesis has been submitted for assessment in partial fulfillment of the PhD degree. The thesis is based on the submitted or published scientific papers which are listed above. Parts of the papers are used directly or indirectly in the extended summary of the thesis. As part of the assessment, co-author statements have been made available to the assessment committee and are also available at the Faculty. The thesis is not in its present form acceptable for open publication but only in limited and closed circulation as copyright may not be ensured.

Preface

This thesis is submitted to the Doctoral School of Engineering and Science at Aalborg University in partial fulfillment of the requirements for the degree of doctor of philosophy. The thesis consists of two parts: an introduction to the research area and a collection of papers that have been published in or submitted to peer-reviewed conferences or journals. The work was carried out in the period from October 2012 to October 2015 at the Department of Architecture, Design, and Media Technology at Aalborg University.

I would like to sincerely thank my supervisor, Prof. Mads Græsbøll Christensen. He gave me the opportunity to work with and explore different ideas on this research topic. His profound knowledge of the field provided me with a very effective supervision. I am very grateful to my co-supervisor Jesper Rindom Jensen who contributed greatly to this work through all of our endless and fruitful discussions.

I would like to thank Prof. Andreas Jakobsson for giving me the opportunity to visit his group at Lund University. I express my gratitude for all our friendly and fruitful discussions on pitch estimation using block sparsity. Also, I am very grateful to Prof. Jacob Benesty who contributed to my work through his ideas in speech enhancement from which I have benefited greatly. A special thanks goes to Nasser Mohammadiha for sharing his expertise on hidden Markov modeling.

I am also grateful to my colleagues at Aalborg University in the past three years. Last, but not least, I thank my friends, family and, most of all, my parents and dear wife Neda for love and support.

Sam Karimian-Azari
Aalborg University, July 6, 2016

Part I

Introduction

Parameter Estimation and Enhancement of Periodic Signals

This thesis considers, first, the application of multichannel signal enhancement of harmonic signals, and second, the estimation of the parameters of such signals from noisy signals. The present introduction is meant as an overview to introduce the main contributions in this thesis, and is organized as follows. The theoretical backgrounds and motivations of the research are represented in order to formulate the problem in section one of the introduction. The second and third sections outline different approaches to solve the problem and include the methodology as well as a number of algorithms used in the following papers. Then, an overview of the contributions of this research is given in Section 4. Finally, the works are summarized and concluded in Section 5.

1 Motivation and Background

Sound is one of the most important elements of multimedia devices. It is also applied in various audio systems such as hearing-aids [77, 112], teleconference systems [67], music information retrieval [15], and diagnosis of illnesses [50, 75]. With the advance of technology, such devices have become increasingly popular in our daily life. Especially mobile devices with powerful processors and a lot of memory lead to changes in consumers' expectations and demands for the best quality of sound. The sound waves are typically converted to electrical signals in order to represent, manipulate, and transform audio signals. Such mobile devices are used in various noisy spaces, and they receive various kinds of noise that have profound impacts on the received signals. For example, a hearing-aid device is used to focus on a particular audio source while canceling out other simultaneous audio

sources in order to increase the intelligibility of the speech signal of interest to a listener. Hence, we wish to enhance the audio signal by reducing noise and other interference. Noise reduction, or speech enhancement, is a difficult task, and it has been a long-term challenging problem in the audio and speech community with the common aim to attenuate the noise as much as possible while the desired signal is unchanged.

1.1 Model-Based Signal Enhancement

Digital signal processing (DSP) concerns signals in digital form and aims to address a variety of challenging problems by numerical algorithms. DSP has evolved dramatically in the field of audio and speech signal processing for performing complex tasks. The choice of numerical algorithms and methods is often reliant on some a priori knowledge of the signal being analyzed in audio compression and coding [80, 81], music analysis and synthesis [68], speech enhancement [18, 76], etc. In this thesis, we aim to enhance audio and speech signals. The algorithms in signal enhancement are commonly achieved by filtering, which is driven either by noise estimates or the signal of interest [59]. For instance, the Wiener filter is designed based on a priori knowledge of the noise statistics that requires an accurate noise estimate even during speech activity [90]. The noise signal is not available directly to estimate it, unless during silence periods of the signal of interest, which is not really practical with dynamic noise. The noise statistics can also be estimated during speech activity [33, 56, 79]. However, the most algorithms for noise estimation do not respond quickly to increasing noise levels [56, 76] and a noisy mixture of multiple speakers. On the other hand, we can reduce the noise concerning a model of the signal of interest, and predict the signal of interest from a number of input noisy signals using the parameters of the model. Therefore, we can design an optimal filter to reconstruct the signal of interest based on the model of the signal.

To the best of our study, the most used speech models are the linear prediction (LP) model [43, 49, 73, 86], the hidden Markov model (HMM) [40, 43, 94], and the harmonic model [25, 59, 64, 83, 84, 87]. The LP model, also called the autoregressive (AR) model, can represent human speech production by modeling the vocal tract as an all-pole system. The excitation to the system is either a random noise (for unvoiced speech) or impulses with the period corresponding to the fundamental frequency of voiced speech [73]. Several methods have been proposed to estimate the parameters of the LP model from the noisy signals. For example, a maximum a posteriori (MAP) estimate of the parameters has been used to estimate a clean speech signal through an optimal Wiener filter in the frequency domain [73]. In addition, the Kalman filtering is another technique used to estimate the clean signal based on the LP model [86] that is extended with the assumption of

time-varing parameters [78] and iteratively optimized for unknown parameters [49]. The HMM-based noise reduction is a robust spectral estimator of the clean signal in two steps: the training step to estimate the probability distribution of the clean signal and noise, and the construction of the desired signal [18]. The harmonic (sinusoidal) model represents *harmonic sounds* as a sum of harmonically related sine waves. Therefore, the noisy signals can be decomposed into the harmonics of the harmonic sounds and noise, and we can refine the filtering techniques with no distortion of the amplitudes of the harmonics [25, 84].

The sounds that may include a number of periodic signals in the background can potentially be separated [83, 87] if their frequency components do not overlap at the same time. Otherwise, we can make use of spatial information of the audio sources to separate them. Some of the audio devices use an array of microphones at known locations to spatially process sound waves. The devices receive temporal samples across the microphones that we commonly name as spatiotemporal samples. These devices can potentially perform spatial filters to align the multichannel signals regarding time delays between the channels [5, 12, 108]. Hence, a very precise estimate of the time-differences-of-arrival (TDOA) is required to design a spatial filter.

1.2 Audio Signal Modeling

Human speech is generated either by vocal cord vibration or open glottis without the vocal cord vibration that leads to turbulent airflow [54]. The vocal cord vibration yields vowels which are quasi-periodic in the time domain and commonly named as voiced speech. The other part that is voiceless (unvoiced) has different fricative sounds such as /s/ and /sh/. These sounds are usually approximated by Gaussian noise to simulate the signal of airflow. Fig. 1 depicts the spectrogram and waveform of the signal of a female speech uttering "first succeeded", including both voiced and unvoiced sounds. By dividing the speech signal into short intervals (frames), the voiced parts have a regular spacing between the significant spectral components in a low frequency range. The frames of the unvoiced speech have high average power in a high frequency range, without a significant structure in the power spectrum. For example, the voiced and unvoiced phonemes /fir/ and /s/ are recognizable respectively at the beginning of the sentence and after a short silence.

We can identify different classes of a speech signal (voiced, unvoiced, and silence) from the spectral information of the signal. Distinguishing between speech presence and silence in an utterance is a crucial task in many applications of speech technology such as speech recognition [56], coding [7], VoIP [95], and enhancement [58]. For example, in the Wiener filter for noise reduction, the noise statistics can be estimated and updated during the silent

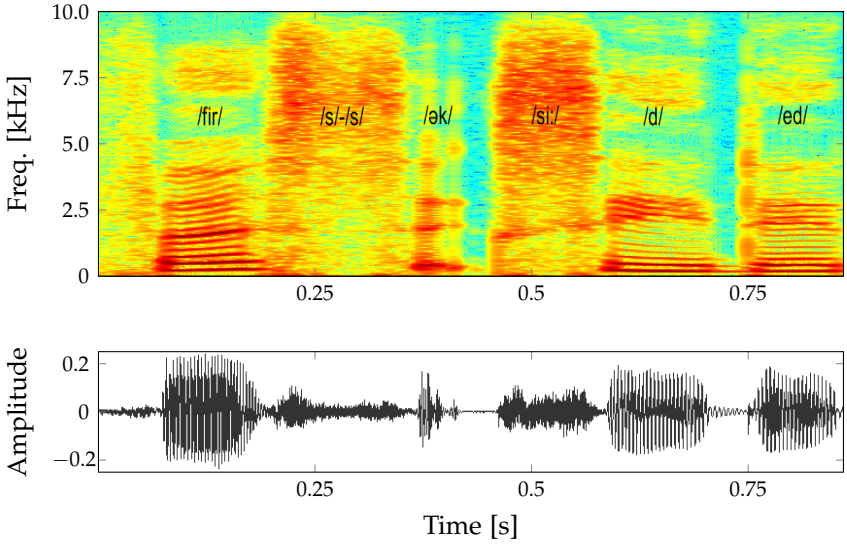


Fig. 1: Spectrogram and waveform of a voiced/unvoiced speech signal.

periods [76]. Different algorithms have been proposed to realize voice activity detection (VAD) that is often not trivial in the presence of noise [90]. As a solution in low signal-to-noise ratio conditions, the harmonic structure of voiced speech can also be used in the VAD [47].

The majority of musical instruments nearly have harmonic sounds [54] with a quasi-periodic waveform. They consist of frequency components at integer multiples of the fundamental frequency, or pitch as it is commonly known. However, for some musical instruments with the inharmonicity property, the harmonics are not exact integers of the fundamental frequency. This imperfection problem can be modeled underlying its physical phenomenon [45] or with perturbed harmonics in a general model [30, 51]. Despite the inharmonicity, the structure of most musical instruments is similar to the spectrum of voiced speech. The example in Fig. 2 shows the spectrogram of a clarinet sound with changing pitch for 3.5 seconds and its waveform in a short frame of 0.1 seconds.

In the following, we continue to formulate signals of harmonic sounds and provide models of single-channel and multichannel signals in the presence of additive noise. We apply the models, more specifically, in model-based signal estimation and enhancement methods.

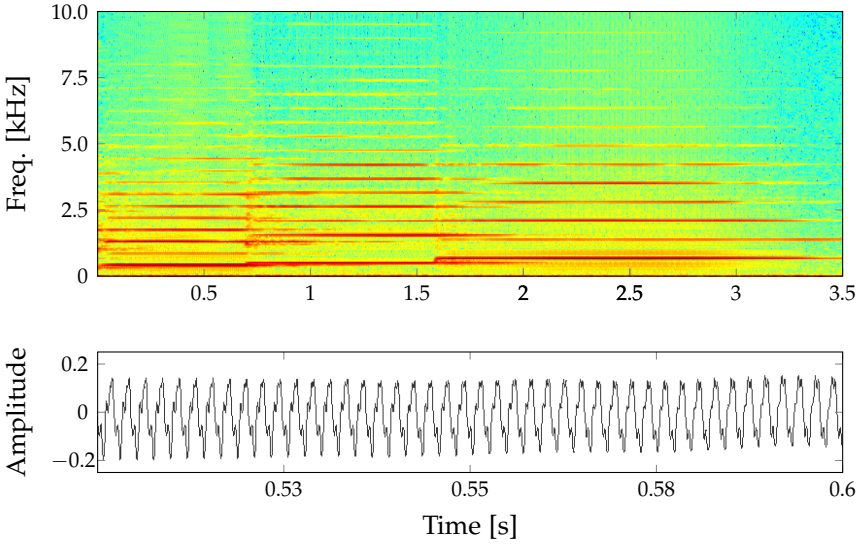


Fig. 2: Spectrogram of a clarinet sound signal and its waveform in 100 ms.

Single-Channel Signal Model

Spectral estimation is generally applied for signal analysis and estimation of the total power distribution over the frequency spectrum [101]. *Non-parametric* spectral estimators are typically suggested without a priori assumption on the signal. The most of the non-parametric estimators are implemented using bandpass filters on uniformly spaced frequencies (e.g., the basic periodogram estimator [101]). This is equivalent to decomposing the signal to the uniformly spaced frequency bands $v_k = 2\pi(k-1)/K$ for $k = 1, 2, \ldots, K$ in the discrete time index n as

$$x(n) = \sum_{k=1}^{K} b_k e^{\jmath v_k n}, \tag{1}$$

where v_k is the normalized frequency in radian-per-second of the basic sinusoidal components at K uniformly spaced frequencies with the complex amplitudes $b_k = |b_k| e^{\jmath \phi_k}$ and phases ϕ_k, and $\jmath = \sqrt{-1}$. This model outlines a structure of an amplitude spectrum and provides a general formulation in the case that an accurate model of the signal is not available. Usually, input signals are subject to a time development, and the parameters of the signal in the time index n may involve the whole set of observations $\mathbf{x}(n) = \left[\begin{array}{cccc} x(n) & x(n-1) & \cdots & x(n-N+1) \end{array} \right]^T$, where the superscript ^{T} is the transpose operator. Observed signals are expressed in the presence of the

additive noise $\mathbf{v}(n)$ as

$$\begin{aligned}\mathbf{y}(n) &= \left[\begin{array}{cccc} y(n) & y(n-1) & \cdots & y(n-N+1) \end{array}\right]^T \\ &= \mathbf{x}(n) + \mathbf{v}(n),\end{aligned} \tag{2}$$

where the unwanted noise is assumed to be a zero-mean random signal and uncorrelated with $\mathbf{x}(n)$. We implicitly assume that the characteristics of the signal and the noise are stationary over that interval. Hence, we can express the input signal vector using the non-parametric signal model in (1) as

$$\mathbf{y}(n) \triangleq \mathbf{Z}_{\mathrm{b}}\mathbf{b}(n) + \mathbf{v}(n), \tag{3}$$

where $\mathbf{Z}_{\mathrm{b}}$ is the Vandermonde matrix of the order N including the discrete Fourier vectors $\mathbf{z}_{\mathrm{t}}(v_k) = \left[\begin{array}{cccc} 1 & e^{-\jmath v_k} & \cdots & e^{-\jmath v_k(N-1)} \end{array}\right]^T$ such that

$$\mathbf{Z}_{\mathrm{b}} = \left[\begin{array}{cccc} \mathbf{z}_{\mathrm{t}}(v_1) & \mathbf{z}_{\mathrm{t}}(v_2) & \cdots & \mathbf{z}_{\mathrm{t}}(v_K) \end{array}\right], \tag{4}$$

and

$$\mathbf{b}(n) = \left[\begin{array}{cccc} b_1 e^{\jmath v_1 n} & b_2 e^{\jmath v_2 n} & \cdots & b_K e^{\jmath v_K n} \end{array}\right]^T. \tag{5}$$

This kind of general signal decomposition in frequency bands can typically be used in spectral estimation [101] and signal enhancement [76] of either voiced or unvoiced sounds. The earliest enhancement method in the frequency domain dates back to the 1960s [97, 98] where the signal of interest is estimated by subtracting an estimate of the noise spectrum from the noisy signal. The first popular algorithm in spectral subtraction was proposed by exploiting the fact that the noise is additive [11]. However, speech signals may be distorted due to an suboptimal averaging of the noise spectrum measured during silent periods. An overestimate of the noise spectrum affords an effect on the intelligibility of speech and introduces an annoying musical noise [8]. To avoid such a problem, some modifications have been done in [8, 99]. Although the spectral subtraction is not derived in an optimal way, we can apply the Wiener filtering approach to attain the minimum error over the spectrum [76].

The examination of spectrogram and waveform has shown that voiced speech and many musical instruments have the *harmonic model*. The concept of model-based spectral estimation is to extract information from some data underlying the model that can formulate the signal. In the harmonic model, the sinusoidal components are harmonically related such that

$$x(n) = \sum_{l=1}^{L} a_l e^{\jmath \omega_l n}, \tag{6}$$

where L is the harmonic model order, $a_l = |a_l|e^{j\psi_l}$ is the complex amplitude of the lth harmonic with the normalized frequency ω_l and phase ψ_l. We compute the discrete-time analytic signal generally to simplify the notation and reduce complexity [31], and real-life signals are mapped to the analytical counterpart using the Hilbert transform and transferred back by taking only the real part of the complex signal [52]. The harmonics are ideally related to the fundamental frequency[1] ω_0 such that the frequencies of the harmonics are defined as integer products of the fundamental frequency, i.e., $\omega_l = l\omega_0$. Spectral estimation of such a harmonic signal is reduced to estimate the parameters of the model which is more accurate than a non-parametric estimator when the model holds. We represent the input signal vector (2) concerning only the spectrum of the harmonics such that

$$\mathbf{y}(n) \triangleq \mathbf{Z}_{\mathrm{t}}\mathbf{a}(n) + \mathbf{v}(n), \tag{7}$$

where

$$\mathbf{Z}_{\mathrm{t}} = \begin{bmatrix} \mathbf{z}_{\mathrm{t}}(\omega_1) & \mathbf{z}_{\mathrm{t}}(\omega_2) & \cdots & \mathbf{z}_{\mathrm{t}}(\omega_L) \end{bmatrix}, \tag{8}$$

and

$$\mathbf{a}(n) = \begin{bmatrix} a_1 e^{j\omega_0 n} & a_2 e^{j2\omega_0 n} & \cdots & a_L e^{jL\omega_0 n} \end{bmatrix}^T. \tag{9}$$

Once the fundamental frequency and the model order of the harmonics are found from the noisy signals, either separately [24, 102] or jointly [23] (they will be explained in the next section), we can estimate the clean signal with respect to (w.r.t.) its model. For example, we can estimate the signal using the model-based filtering technique [25] and the expectation maximization (EM) algorithm that is a maximum likelihood (ML) estimator involving the parameter estimates [24, 44].

Multichannel Signal Model

We reformulate the signal model for multichannel signals. For an array that includes M omnidirectional microphones, the observed signal at the mth microphone ($m = 1, 2, \ldots, M$) is given by

$$x_m(n) = x(n - f_{\mathrm{s}}\tau_m), \tag{10}$$

where f_{s} is the sampling frequency, and τ_m is the relative delay between the received signal at the mth and the first microphone. In the remainder of this thesis, we assume a uniform linear array (ULA) that is easily extendable to other array structures. For a far-field setup of the array respective to the

[1] The fundamental frequency of voiced speech is about 60–150 Hz for male speakers and 200–400 Hz for female and child speakers [88].

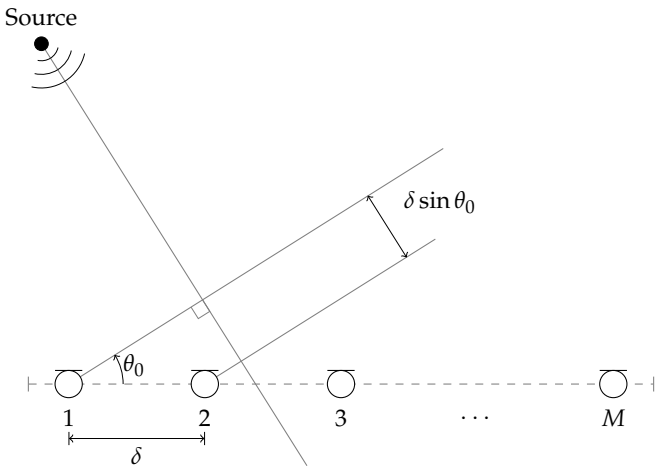


Fig. 3: Uniform linear array of M microphones [61].

audio source at the direction of arrival (DOA) θ_0, illustrated in Fig. 3, the relative delay is therefore given by

$$\tau_m = (m-1)\frac{\delta \sin\theta_0}{c}, \tag{11}$$

where δ is the distance between two successive microphones and c is the speed of the sound. The complex amplitudes of the harmonics are assumed to be approximately identical across the channels for the far-field setup. In a noisy and anechoic environment, the accumulated $M \times N$ spatiotemporal samples,

$$y_m(n) = x_m(n) + v_m(n), \tag{12}$$

are therefore given in a vector form w.r.t. the non-parametric signal model in (1) as

$$\mathbf{y}_{\mathrm{st}}(n) = \mathbf{Z}_{\mathrm{sb}}\mathbf{b}(n) + \mathbf{v}_{\mathrm{st}}, \tag{13}$$

where

$$\mathbf{Z}_{\mathrm{sb}} = \left[\begin{array}{cccc} \mathbf{z}_{\mathrm{st}}(\theta_0, v_1) & \mathbf{z}_{\mathrm{st}}(\theta_0, v_2) & \cdots & \mathbf{z}_{\mathrm{st}}(\theta_0, v_K) \end{array}\right], \tag{14}$$

$$\mathbf{z}_{\mathrm{st}}(\theta_0, v_k) = \mathbf{z}_{\mathrm{s}}(v_k) \otimes \mathbf{z}_{\mathrm{t}}(v_k), \tag{15}$$

and $\otimes$ denotes the Kronecker product of two vectors: $\mathbf{z}_{\mathrm{t}}(v_k)$ and $\mathbf{z}_{\mathrm{s}}(v_k) = \left[\begin{array}{cccc} 1 & e^{-\jmath f_{\mathrm{s}} \tau_m v_k} & \cdots & e^{-\jmath f_{\mathrm{s}} \tau_m v_k (M-1)} \end{array}\right]^T$ that is the discrete Fourier vector of

the linear spatial samples (also known as the steering vector). The vector of the noisy signals is represented w.r.t. to the harmonic model such that

$$\mathbf{y}_{\mathrm{st}}(n) = \mathbf{Z}_{\mathrm{st}}\mathbf{a}(n) + \mathbf{v}_{\mathrm{st}}, \tag{16}$$

where

$$\mathbf{Z}_{\mathrm{st}} = \begin{bmatrix} \mathbf{z}_{\mathrm{st}}(\theta_0, \omega_1) & \mathbf{z}_{\mathrm{st}}(\theta_0, \omega_2) & \cdots & \mathbf{z}_{\mathrm{st}}(\theta_0, \omega_L) \end{bmatrix}. \tag{17}$$

With regards to the linear signal formulation (16), we can estimate the spectral amplitudes $\mathbf{a}(n)$ by exploiting the basis matrix $\mathbf{Z}_{\mathrm{st}}$, or $\mathbf{Z}_{\mathrm{t}}$ for single-channel signals from the given frequencies and DOA of the harmonics.

1.3 Problem Statement

By preserving the spectral information of the signal of interest, we can avoid any degradation in speech intelligibility in the presence of noise and interfering signals [76]. Hence, this research is motivated by the belief that the prior spectral information of the signal of interest can perform speech enhancement methods. This PhD thesis aims at the estimation of the spectral information and enhancement of periodic signals associated with the advantage of their model. We consider the problem for harmonic signals by exploiting the relevant parameters of multichannel signals. In order to conduct such a model-based signal enhancement, we generally require an accurate estimate of the corresponding parameters, i.e.,

$$\boldsymbol{\eta} = \begin{bmatrix} \theta_0 & \boldsymbol{\eta}_0 \end{bmatrix}^T = \begin{bmatrix} \theta_0 & \omega_0 & |a_1| & \psi_1 & \cdots & |a_L| & \psi_L \end{bmatrix}^T. \tag{18}$$

Estimation of these parameters is an issue in the presence of real-life noise, and we focus on a number of methods to estimate the fundamental frequency and the DOA.

2 Harmonic Model-Based Signal Enhancement

A tremendous amount of research has been devoted to audio and speech enhancement as it is a key task for many audio processing applications, see [6, 76] and the references therein. Numerous single-channel methods have been proposed which can be extended to multichannel signals. Most approaches have been classified into three categories [18]: spectral restoration [41], filtering [25, 42, 57, 84], and model-based methods [25, 43, 49, 84, 86]. Spectral restoration technique minimizes the noise by estimating the spectrum of the clean signal, and the filtering technique passes the noisy signal through a linear filter to estimate the signal in the time or the frequency

domain. Both the spectral restoration and the filtering techniques can be designed associated with the models of the clean signal. In other words, such model-based spectral restoration and filtering methods are designed subject to no distortion on the signal. This section briefly summarizes a number of harmonic model-based techniques for spectral restoration of a noisy signal which are directly related to the proposed methods in this thesis.

2.1 Single-Channel Signal Enhancement

Filtering is the most fundamental method of noise reduction that can be formulated in time and frequency domains. We can estimate the signal of interest by passing the noisy signal $\mathbf{y}(n)$ through the filter $\mathbf{h}$ in the time domain such that

$$\hat{x}(n) = \mathbf{h}^H \mathbf{y}(n) \tag{19}$$

$$= \mathbf{h}^H \mathbf{x}(n) + \mathbf{h}^H \mathbf{v}(n). \tag{20}$$

The general principle is to minimize the noise as much as possible. The maximum signal-to-noise ratio (SNR) filter [4] is designed by maximizing the defined output SNR (oSNR) such that

$$\mathbf{h}_{\max} = \arg\max_{\mathbf{h}} \text{oSNR}(\mathbf{h}) \tag{21}$$

$$= \arg\max_{\mathbf{h}} \frac{\mathbf{h}^H \mathbf{R}_{\mathbf{x}} \mathbf{h}}{\mathbf{h}^H \mathbf{R}_{\mathbf{v}} \mathbf{h}}, \tag{22}$$

where the covariance matrices of the clean signal and the noise, respectively, are $\mathbf{R}_{\mathbf{x}} = E\left[\mathbf{x}(n)\mathbf{x}^H(n)\right]$ and $\mathbf{R}_{\mathbf{v}} = E\left[\mathbf{v}(n)\mathbf{v}^H(n)\right]$, and $E[\cdot]$ is the mathematical expectation. The Wiener filter is the well-known solution in noise reduction that restores the desired signal by promising a minimum mean square error (MMSE) of the estimated signal through the filter, i.e.,

$$\mathbf{h}_{\text{W}} = \arg\min_{\mathbf{h}} E\left[|\hat{x}(n) - x(n)|^2\right]. \tag{23}$$

Even though an optimal Wiener filter is noise dependent analytically and improves the oSNR, its output suffers from speech distortion [19]. Regarding the spectrum of the signal, we can intuitively design filter banks having unit gains at the frequencies of the harmonics, i.e., ω_l for $l = 1, 2, \ldots, L$. Therefore, the desired signal is passed undistorted using a bank of L filters at those frequencies such that $\mathbf{H}(\omega_0) = \begin{bmatrix} \mathbf{h}(\omega_1) & \mathbf{h}(\omega_2) & \cdots & \mathbf{h}(\omega_L) \end{bmatrix}$. For example, the comb filter whose frequency response contains peaks at the harmonics [84]. The Capon spectral estimator [16] (also known as the minimum variance distortionless response [MVDR] filter) is an optimal solution with

a minimum noise in the output subject to the distortionless constraint at the frequencies of the harmonics such that

$$\min_{\mathbf{h}(\omega_l)} \mathbf{h}^H(\omega_l)\mathbf{R}_\mathbf{y}\mathbf{h}(\omega_l) \quad \text{subject to} \quad \mathbf{h}^H(\omega_l)\mathbf{z}_\mathrm{t}(\omega_l) = 1, \tag{24}$$

where $\mathbf{R}_\mathbf{y} = E\left[\mathbf{y}(n)\mathbf{y}^H(n)\right]$ is the covariance matrix of the noisy signal. The MVDR filter is given by

$$\mathbf{h}_\mathrm{MVDR}(\omega_l) = \mathbf{R}_\mathbf{y}^{-1}\mathbf{z}_\mathrm{t}(\omega_l)\left[\mathbf{z}_\mathrm{t}^H(\omega_l)\mathbf{R}_\mathbf{y}^{-1}\mathbf{z}_\mathrm{t}(\omega_l)\right]^{-1}. \tag{25}$$

In [23, 25], the filter bank has been integrated into a single filter to pass all the harmonics while minimizing the power at other frequencies. The filter is adaptive with strong background noise reduction and guarantees that the signal of interest is undistorted. This approach is represented as the following optimization problem:

$$\min_{\mathbf{h}(\mathbf{\Omega})} \mathbf{h}^H(\mathbf{\Omega})\mathbf{R}_\mathbf{y}\mathbf{h}(\mathbf{\Omega}) \quad \text{subject to} \quad \mathbf{h}^H(\mathbf{\Omega})\mathbf{Z}_\mathrm{t} = \mathbf{1}^T, \tag{26}$$

where $\mathbf{\Omega} \triangleq \begin{bmatrix}\omega_0 & 2\omega_0 & \dots & L\omega_0\end{bmatrix}^T$, and $\mathbf{1}$ is defined as the all-ones column vector of length L. This problem resembles the known linearly constrained minimum variance (LCMV) filter and it is given by

$$\mathbf{h}_\mathrm{LCMV}(\mathbf{\Omega}) = \mathbf{R}_\mathbf{y}^{-1}\mathbf{Z}_\mathrm{t}\left[\mathbf{Z}_\mathrm{t}^H\mathbf{R}_\mathbf{y}^{-1}\mathbf{Z}_\mathrm{t}\right]^{-1}\mathbf{1}. \tag{27}$$

In practice, the covariance matrix $\mathbf{R}_\mathbf{y}$ is estimated via averaging of the observed signals over time. The statistics of the signal and the noise are assumed to be stationary, which limits the performance of the covariance matrix estimate for the limited number of samples. Moreover, the length of the filter limits the number of samples. An iterative adaptive approach (IAA) has been proposed to estimate the covariance matrix [39, 114]. This approach enables us to estimate the covariance matrix full rank from only a few samples at the expense of an increased computation complexity.

2.2 Multichannel Signal Enhancement

Array processing techniques have been initially developed for applications of narrowband signals such as telecommunication, radar, and sonar [107]. Conventional array processing techniques typically exploit the assumption that the desired signal and interferers are physically separated in space. A beamformer, i.e., a spatial filter, is designed to steer the beam of the array in one direction to estimate the signal with minimum noise in the output. Over the years, many different beamformers have been developed, including both data-independent (fixed) and data-dependent (adaptive) beamformers. In

principle, data-dependent beamformers should yield better results than data-independent beamformers in most scenarios. They can adapt to the acoustic environment, but it is often difficult to estimate the statistics of the noise efficiently and the location of the signal sources accurately which actually causes signal cancellation. Some well-known examples of beamformer designs are the delay-and-sum (DS) [108], maximum SNR [5], multiple sidelobe canceler (MSC) [3], generalized sidelobe canceler (GSC) [14, 48], superdirective [34], minimum variance distortionless reponse (MVDR) [16], and linearly constrained minimum variance (LCMV) [46] beamformers. An overview with more details about various beamformer designs for microphone arrays have been presented in [5, 12, 108] and the references therein.

Broadband beamformers are typically designed by K narrowband beamformers at nonparametric frequency bands v_k, for $k = 1, 2, \ldots, K$, and the DOA θ_0 such that

$$\hat{\mathbf{x}}(n) = \mathbf{h}^H(\theta_0, v_k)\mathbf{y}_{\text{st}}(n). \tag{28}$$

The MVDR beamformer is designed to minimize the output power subject to distortionless constraint on the given DOA at each frequency band, i.e.,

$$\min_{\mathbf{h}(\theta_0, v_k)} \mathbf{h}^H(\theta_0, v_k)\mathbf{R}_{\mathbf{y}_{\text{st}}}\mathbf{h}(\theta_0, v_k) \quad \text{subject to} \quad \mathbf{h}^H(\theta_0, v_k)\mathbf{z}(\theta_0, v_k) = 1, \tag{29}$$

where $\mathbf{R}_{\mathbf{y}_{\text{st}}} = E\left[\mathbf{y}_{\text{st}}(n)\mathbf{y}_{\text{st}}^H(n)\right]$ is the covariance matrix of the noisy spatiotemporal signals. The covariance matrix is estimated via either averaging of the signals over an interval, or the IAA that has been introduced as the recursive least squares (RLS) algorithm in [107] (e.g., [60]). The MVDR beamformer is given by

$$\mathbf{h}_{\text{MVDR}}(\theta_0, v_k) = \mathbf{R}_{\mathbf{y}_{\text{st}}}^{-1}\mathbf{z}(\theta_0, v_k)\left[\mathbf{z}^H(\theta_0, v_k)\mathbf{R}_{\mathbf{y}_{\text{st}}}^{-1}\mathbf{z}(\theta_0, v_k)\right]^{-1}, \tag{30}$$

and the DS beamformer is given as a special case of the MVDR beamformer in white noise as

$$\mathbf{h}_{\text{DS}}(\theta_0, v_k) = \mathbf{z}(\theta_0, v_k)\left[\mathbf{z}^H(\theta_0, v_k)\mathbf{z}(\theta_0, v_k)\right]^{-1}. \tag{31}$$

A fine spectral estimate result in small frequency grids with the cost of an increased computational complexity. Although, the distortionless constraint may not be valid of the true harmonics in a case with large frequency grids. Moreover, broadband beamformers may pass non-coherent noise at frequency bands other than the frequencies of the desired signal. Use of the spectral information is an approach for speaker separation [87], and in conjunction with a microphone array, beamforming provides a versatile tool in separation of audio sources concerning their DOAs [108]. Hence,

the harmonic-model filtering approach in (27) can be extended into the spatiotemporal filtering with the following problem formulation [63]:

$$\min_{\mathbf{h}(\theta_0,\mathbf{\Omega})} \mathbf{h}^H(\theta_0,\mathbf{\Omega})\mathbf{R}_{\mathbf{y}_{\text{st}}}\mathbf{h}(\theta_0,\mathbf{\Omega}) \quad \text{subject to} \quad \mathbf{h}^H(\theta_0,\mathbf{\Omega})\mathbf{Z}_{\text{st}} = \mathbf{1}^T. \tag{32}$$

The solution is given by

$$\mathbf{h}_{\text{LCMV}}(\theta_0,\mathbf{\Omega}) = \mathbf{R}_{\mathbf{y}_{\text{st}}}^{-1}\mathbf{Z}_{\text{st}}\left[\mathbf{Z}_{\text{st}}^H\mathbf{R}_{\mathbf{y}_{\text{st}}}^{-1}\mathbf{Z}_{\text{st}}\right]^{-1}\mathbf{1}. \tag{33}$$

Acknowledging the importance of the use of a priori knowledge about the spectrum of the signal, we present a class of parametric beamformers in paper A. The beamformers achieve better results in noise reduction with a minimum noise in the frequencies between the harmonics, and improve speech intelligibility in comparison with the non-parametric beamformers.

2.3 Performance Measures

We usually apply some criterion to evaluate the performance of speech enhancement. In developing the algorithms for speech enhancement, the subjective is to improve intelligibility and the quality of speech signals. In order to verify such algorithms, different performance criterion are utilized: subjective listening tests and objective measures. The subjective listening test provides quality and intelligibility evaluations that involve comparison of the original and enhanced signal by some listeners [76] (e.g., mean-opinion-score tests [69]). In most cases, subjective listening tests are indeed time-consuming to train the listeners and conduct the test. In contrast, objective evaluations involve some measures between the original and the enhanced signal. Some commonly used objective measures include signal-to-noise ratio (SNR), speech distortion index [19], perceptual evaluation of speech quality (PESQ) measure [92], and the recently proposed measurement of short-time objective intelligibility measure (STOI) [104]. These mathematical measures can also be integrated by some algorithms to optimize them, e.g., the maximum SNR filter in paper A.

The speech distortion index is given by the ratio of the power of the difference between the filtered and unfiltered desired signals to the power of the desired signal as

$$v_{\text{sd}}(\mathbf{h}) = \frac{E\left[\left|\mathbf{h}^H\mathbf{x}(n) - x_1(n)\right|^2\right]}{E\left[\left|x_1(n)\right|^2\right]}. \tag{34}$$

The LCMV filters in (27) and (33), which are designed to be subjected to no distortion on the harmonics, gives $v_{\text{sd}}(\mathbf{h}_{\text{LCMV}}) = 0$. Paper A also proposes

a trade-off filter, $\mathbf{h}_{\mathrm{T}}$. The performance of the filter is between the maximum SNR and the LCMV filters such that

$$\mathrm{oSNR}\left(\mathbf{h}_{\max}\right) \geq \mathrm{oSNR}\left(\mathbf{h}_{\mathrm{T}}\right) \geq \mathrm{oSNR}\left(\mathbf{h}_{\mathrm{LCMV}}\right) \tag{35}$$

and

$$v_{\mathrm{sd}}\left(\mathbf{h}_{\max}\right) \geq v_{\mathrm{sd}}\left(\mathbf{h}_{\mathrm{T}}\right) \geq v_{\mathrm{sd}}\left(\mathbf{h}_{\mathrm{LCMV}}\right) = 0. \tag{36}$$

Such performance measures show that the noise reduction with the maximum output SNR is not necessarily optimal when the output signal is distorted.

3 Fundamental Frequency and Direction-of-Arrival Estimation

The performance of the filters, which are designed based on the model of the signal of interest, degrades, if the parameter estimates of the model are inaccurate. For example, the output SNR of beamformers degrades in the presence of steering vector errors due to direction-of-arrival (DOA) and frequency errors. Although a number of methods have been proposed to lessen this problem [35, 72], we introduce optimal solutions to estimate the parameters of the model with the minimum error.

In estimation theory, the measured signals are assumed to be stochastic with a probability distribution depending on the noise and the parameters of the model, and, for a sufficiently large number of samples, the probability density function of the measured signals has a normal distribution around an expectation according to the central limit theory [10]. Most parameter estimation methods rely on the assumption that the signal of interest is stationary over a set of samples. Although the parameters of speech signals are dynamic, we can assume that the speech signals are quasi-stationary over short intervals, about 20–30 ms, which consequently limits the number of samples and the accuracy of the estimates [101]. Some efforts have recently been made for non-stationary signals using a linear chirp model of increasing/decreasing frequency over time in order to estimate the related parameters [32, 38, 103].

The accuracy of the parameter estimates is an issue in the presence of noise. The noise is usually considered as a stochastic signal with characteristics that limit the accuracy of the estimates of deterministic parameters. Real-life signals are captured in different situations, e.g., exhibition halls, restaurants, streets, airports and train stations, in the presence of noise sources such as a crowd of people (babble noise) and cars. The long-term average of the power spectrum of the noise signals in real-life is different from white

noise [65] that is commonly assumed in some noise reduction [53, 73] and parameter estimation [24] algorithms. Hence, a noise robust estimator is required in real-life applications.

Many non-parametric methods have been proposed to estimate the fundamental frequency [55], which are essentially based on the similarities of the observations, for example, the auto-correlation based method [89]. In general, the non-parametric methods do not have a unique estimate [24]. Another class of methods are parametric which are devised from classical spectral estimation approaches [101]. The estimation of the fundamental frequency, as the parameter of the deterministic signal model, has been investigated in [24] and is generally categorized in three groups:

- statistical methods
- filtering methods
- subspace methods

For the deterministic parameters, deviation of the observations cannot be known exactly, but it is possible to make probabilistic statements from statistics of the observations. In statistical methods, the probability of the parameter of interest is maximized to find the maximum likelihood (ML) estimate [61, 62, 71, 85], or the maximum a posteriori (MAP) estimate [51, 105] that is obtained from Bayesian statistics. Although the statistical methods are statistically efficient and attain the Cramér-Rao lower bound (CRLB) for a high number of samples, the statistical methods are typically restricted to the white Gaussian noise assumption [24]. The filtering methods, as another approach, are derived intuitively based on the concept of the harmonic model-based filtering for signal enhancement [25, 61, 63]. However, the filtering methods do not attain the CRLB [29]. The subspace methods are based on the principle in the Euclidean space that the vector space of the noisy signal includes a subspace of the clean signal and a subspace of the noise. Some fundamental frequency estimators have been proposed in [27, 30, 115] based on the subspace orthogonality of the harmonics. The most common approaches are based on the eigen decomposition of the orthogonal subspaces, e.g., the multiple signal classification (MUSIC) [96] and the estimation of signal parameters by rotational invariance techniques (ESPRIT) [93], which are computationally complex with a biased fundamental frequency estimate [24, 26].

Audio source localization, i.e., DOA estimation, is necessary for an audio system with microphone arrays, and it is a challenging problem in the presence of noise and other interfering sources [13]. The existing localization techniques are generally defined for non-parametric signals, which may be divided into two categories:

- The steered response power (SRP) of a non-parametric beamformer, which scans various locations and searches for a peak in the output of

the beamformer to estimate the DOA of a signal source [36] or multiple sources [111].

- The time-differences-of-arrival (TDOA) estimation [21], e.g., the generalized cross-correlation of the phase transform (GCC-PHAT) between two received signals [70] and the least squares estimator [17, 20].

The performance of localization methods generally depends on the level of the noise, the quantity of employing microphones, and the spacing between the microphones. Moreover, the performance of the SRP-based methods depends on the performance of the applying beamformer. Conventional beamformers are highly frequency-dependent, and their steered response is proportional to increasing frequency [22] that limits the performance of the SRP-based methods. A number of frequency invariant beamformers have been proposed in [74, 110] to perform the SRP-based methods [109]. The high-resolution spectral analysis based on the spatiospectral correlation matrix estimate is a modification of the SRP methods that results in sharp peaks with a high resolution [93, 96]. The cost of this approach is high computation complexity, and in the presence of reverberation, the noise and the source are highly correlated which leads to the removal of the signal of interest as well as the noise [36]. The TDOA-based estimators are preferable because of a computational advantage over the SRP-based methods. The TDOA-based methods are commonly limited on a single-source with poor results in noise. However, we can extend the TDOA-based estimators to scenarios with multiple sources using the model of harmonic signals, e.g., the position-pitch plane based (POPI) estimator [113], and increase the accuracy concerning the noise statistics, e.g., the statistically efficient DOA estimator in [62]. We can estimate the DOA in the same fashion as in the estimation of the fundamental frequency. Moreover, joint DOA and fundamental frequency estimation can be concluded using the nonlinear least squares (NLS) [61], the spatiotemporal filtering [63], and the subspace [115] methods. In multiple source scenarios, the estimation of the fundamental frequency and the DOA jointly has advantages over separate estimation in situations with overlapping DOAs or fundamental frequencies, as long as the other one is distinct.

3.1 Filtering Methods

The harmonic model-based filters for speech enhancement have the minimum output power while they pass the desired signal undistorted. We can also use such filters for two-dimensional (2D) spectral estimation [82]. The parameter estimates of the model can then be obtained by maximizing the output power of the filters, i.e., $E\left[\left|\hat{x}(n)\right|^2\right]$. We can use either the designed

filter bank $\mathbf{H}(\omega_0)$ such that

$$\{\hat{\omega}_0\}_{\mathbf{H}(\omega_0)} = \arg\max_{\omega_0} \operatorname{tr}\left\{\mathbf{H}^H(\omega_0)\mathbf{R}_\mathbf{y}\mathbf{H}(\omega_0)\right\}, \tag{37}$$

or the designed LCMV filter $\mathbf{h}_{\text{LCMV}}(\mathbf{\Omega})$ such that [28]

$$\{\hat{\omega}_0\}_{\mathbf{h}_{\text{LCMV}}(\mathbf{\Omega})} = \arg\max_{\omega_0} \mathbf{h}_{\text{LCMV}}^H(\mathbf{\Omega})\mathbf{R}_\mathbf{y}\mathbf{h}_{\text{LCMV}}(\mathbf{\Omega}) \tag{38}$$

$$= \arg\max_{\omega_0} \mathbf{1}^T \left[\mathbf{Z}_\text{t}^H\mathbf{R}_\mathbf{y}^{-1}\mathbf{Z}_\text{t}\right]^{-1}\mathbf{1}, \tag{39}$$

where $\operatorname{tr}(\cdot)$ denotes the trace of a square matrix. The filter bank has cross-terms which do not appear in (37), and the result of the filter bank is not the same as the result of the LCMV filter [24]. The filtering solution has also been extended to estimate the fundamental frequency and the DOA jointly in [63] using the spatiotemporal filter $\mathbf{h}_{\text{LCMV}}(\theta_0, \mathbf{\Omega})$ such that

$$\{\hat{\theta}_0, \hat{\omega}_0\}_{\mathbf{h}_{\text{LCMV}}(\theta_0,\mathbf{\Omega})} = \arg\max_{\{\theta_0,\omega_0\}} \mathbf{h}_{\text{LCMV}}^H(\theta_0, \mathbf{\Omega})\mathbf{R}_{\mathbf{y}_\text{st}}\mathbf{h}_{\text{LCMV}}(\theta_0, \mathbf{\Omega})$$

$$= \arg\max_{\{\theta_0,\omega_0\}} \mathbf{1}^T \left[\mathbf{Z}_\text{st}^H\mathbf{R}_{\mathbf{y}_\text{st}}^{-1}\mathbf{Z}_\text{st}\right]^{-1}\mathbf{1}. \tag{40}$$

Although the resulting filters are data-dependent, the mean squared error (MSE) of the estimates does not reach the CRLB [63]. Moreover, these estimators require matrix inversions and products for each point in the search grid, which are computationally expensive.

3.2 Statistical Methods

An optimal estimator is associated with the deterministic signal model to find the most likely probability density function (PDF) of the signal or the parameters of the corresponding model. In general, an ML estimator is associated with the given parameters by maximizing the PDF of the signal. The least squares (LS) estimator is the most famous solution with statistically efficient results in white Gaussian noise. In order to estimate the fundamental frequency and the DOA by fitting the given data in the LS estimator, the nonlinear least squares (NLS) estimator is represented by

$$\{\hat{\theta}_0, \hat{\omega}_0\}_\text{NLS} = \arg\min_{\{\theta_0,\omega_0\}} \|\mathbf{y}_\text{st}(n) - \mathbf{Z}_\text{st}\mathbf{a}(n)\|_2^2. \tag{41}$$

The NLS method has the solution for joint fundamental frequency and DOA estimates in white Gaussian noise [61] given by

$$\{\hat{\theta}_0, \hat{\omega}_0\}_\text{NLS} = \arg\max_{\{\theta_0,\omega_0\}} \mathbf{y}_\text{st}^H(n)\mathbf{Z}_\text{st}(\mathbf{Z}_\text{st}^H\mathbf{Z}_\text{st})^{-1}\mathbf{Z}_\text{st}^H\mathbf{y}_\text{st}(n). \tag{42}$$

The NLS estimator is statistically efficient in white noise, i.e., it has the lowest possible variance of the estimate. We can achieve an approximate NLS (aNLS) estimator when $N \times M \to \infty$ such that [24, 61]

$$\{\hat{\theta}_0, \hat{\omega}_0\}_{\text{aNLS}} = \arg \max_{\{\theta_0, \omega_0\}} \left\| \mathbf{y}_{\text{st}}^H(n) \mathbf{Z}_{\text{st}} \right\|_2^2 \tag{43}$$

$$= \arg \max_{\{\theta_0, \omega_0\}} \sum_{l=1}^{L} \left| \mathbf{y}_{\text{st}}^H(n) \mathbf{z}(\theta_0, l\omega_0) \right|^2. \tag{44}$$

The resulting joint DOA and fundamental frequency estimator is computationally simpler than the NLS. The estimates are obtained by locating the highest peak in the sum of the power spectrum of the harmonics. The aNLS is the same as the harmonic summation (HS) solution [85] in the fundamental frequency estimation that fits harmonic sine-waves to the input data. In papers E and F, we apply the harmonic summation to extend the SRP method in order to estimate the parameters of harmonic signals.

In the following, we introduce two estimators of the fundamental frequency and the DOA in a sequential process, and extend them into a solution of joint estimates in paper B. A statistically efficient solution has been proposed in [71] to estimate the fundamental frequency from the location of spectral peaks of the harmonics, $\hat{\omega}_l$, which we call unconstrained frequency estimates (UFEs), and the corresponding power spectrum estimates $|\hat{a}_l|^2$. The estimator is based on the weighted least-squares (WLS) solution. The weighting matrix of the WLS approach is given by the corresponding Fisher information matrix (FIM) under the white Gaussian noise assumption. The estimator is the sum of the weighted UFEs such that

$$\{\hat{\omega}_0\}_{\text{WLS}} = \sum_{l=1}^{L} l \, |\hat{a}_l|^2 \, \hat{\omega}_l / \sum_{l=1}^{L} l^2 \, |\hat{a}_l|^2. \tag{45}$$

A DOA estimator has been proposed in [62] based on mutual coupling of the multichannel phase estimates of the given harmonics. The estimator consists of two steps using the WLS method that the weighting matrices are given by the FIM in the assumption of white Gaussian noise. The last step of the WLS DOA estimator is designed w.r.t. the DOA estimates $\hat{\theta}_l$ of the given harmonics as

$$\{\hat{\theta}_0\}_{\text{WLS}} = \sum_{l=1}^{L} l^2 \, |\hat{a}_l|^2 \, \hat{\theta}_l / \sum_{l=1}^{L} l^2 \, |\hat{a}_l|^2. \tag{46}$$

The WLS estimators are computationally simpler than the corresponding NLS estimators, though they achieve similar performance to the NLS method for a large number of samples and/or high SNR [71].

To avoid the white noise assumption in the aforementioned WLS estimators, we have shown in paper B that the additive noise can be converted into an equivalent phase-noise in each sinusoid [106]. We can therefore approximate the multichannel noisy signal as the sum of harmonics with the phase-noise $\Delta\psi_{l,m}(n)$ for the lth harmonic at the time instance n and the microphone m, i.e.,

$$\begin{aligned} y_m(n) &= x_m(n) + v_m(n) \\ &\approx \sum_{l=1}^{L} a_l e^{j\left(\omega_l n - \omega_l f_s \tau_m + \Delta\psi_{l,m}(n)\right)}. \end{aligned} \tag{47}$$

We therefore model the UFEs and the DOA estimates of the harmonics with multivariate normal distributions, respectively, such as

$$\hat{\boldsymbol{\Omega}} = \begin{bmatrix} \hat{\omega}_1 & \hat{\omega}_2 & \dots & \hat{\omega}_L \end{bmatrix}^T \triangleq \mathcal{N}\left(\mathbf{d}_L \omega_0, \mathbf{R}_{\Delta\boldsymbol{\Omega}}\right), \tag{48}$$

$$\hat{\boldsymbol{\Theta}} = \begin{bmatrix} \hat{\theta}_1 & \hat{\theta}_2 & \dots & \hat{\theta}_L \end{bmatrix}^T \triangleq \mathcal{N}\left(\mathbf{1}\theta_0, \mathbf{R}_{\Delta\boldsymbol{\Theta}}\right), \tag{49}$$

where $\mathcal{N}(\cdot,\cdot)$ denotes the normal distribution that its first argument is the vector of the expected values and the second one is the covariance matrix of the variables, and $\mathbf{d}_L = \begin{bmatrix} 1 & 2 & \dots & L \end{bmatrix}^T$. The covariance matrices $\mathbf{R}_{\Delta\boldsymbol{\Omega}}$ and $\mathbf{R}_{\Delta\boldsymbol{\Theta}}$ are related to the diagonal matrix $\mathbf{R}_{\Delta\boldsymbol{\Psi}}$ which includes the reciprocal narrowband SNRs of the harmonics (i.e., $\Phi(\omega_l)/|a_l|^2$) such that

$$\mathbf{R}_{\Delta\boldsymbol{\Psi}} = \frac{1}{2}\,\mathrm{diag}\left\{ \frac{\Phi(\omega_1)}{|a_1|^2} \;\; \frac{\Phi(\omega_2)}{|a_2|^2} \;\; \dots \;\; \frac{\Phi(\omega_L)}{|a_L|^2} \right\}, \tag{50}$$

where $\Phi(\omega_l)$ is the narrowband power spectrum of the noise at the lth harmonic. Therefore, the covariance matrices can be estimated in practice from statistics of the UFEs and the DOA estimates of the harmonics. We propose then the fundamental frequency and the DOA estimators using the spectral characteristics of the noise and show that the estimators are robust against colored Gaussian noise.

The number of harmonics, L, must be known for the aforementioned methods, though the estimation of the number of harmonics is a difficult problem known as model order estimation. The maximum a posteriori (MAP) [37] (see also [102]), the Akaike information criterion (AIC) [2], and the minimum description length criterion (MDL) [91] are the most common solutions for the model order selection. The MAP is commonly used for random parameters whose a priori density function is known, and it is defined from the FIM that results as a penalty term. Hence, the MAP model order estimator is presented by the penalized MSE of an ML estimator: either the LS amplitude estimator [102] or the NLS (or the optimal model-based filter) fundamental frequency estimator [23] that obtains joint estimates of the fundamental frequency and the model order. No similar model order estimation has been

considered for the multichannel case before. In the following, we incorporate the statistical methods (the harmonic summation and the MAP model order estimation) with the non-parametric filter $\mathbf{h}(\theta_0, v_k)$ to estimate the model order of periodic signals from multichannel signals in papers E and F.

In a scenario with multiple harmonic sources, the estimation problem turns into a multi-pitch estimation problem. This problem can be solved by extending most of the fundamental frequency estimation solutions given a priori knowledge of the number of measured sources as well as the model order of the harmonic sources [24]. Some methods have been proposed to avoid such assumptions; for example, the non-negative matrix factorization (NMF) method [9, 100] that decomposes the spectrogram of the signal into two matrices to represent it in terms of a set of a magnitude spectrum and an activation matrix. A recently proposed pitch estimation using block sparsity (PEBS) technique [1] uses a sparse dictionary. To avoid the assumptions on the number of sources and the model orders, it imposes the large dictionary $\mathbf{W} = \begin{bmatrix} \mathbf{Z}_{\mathrm{t},1} & \mathbf{Z}_{\mathrm{t},2} & \dots & \mathbf{Z}_{\mathrm{t},S} \end{bmatrix}$ on S feasible fundamental frequencies and the maximum harmonics given by the basis matrices $\mathbf{Z}_{\mathrm{t},r}$ for $r = 1, 2, \dots, S$. The PEBS method has two steps: first, sparse spectral amplitudes $\mathbf{a}(n) = \begin{bmatrix} \mathbf{a}_1^T(n) & \mathbf{a}_2^T(n) & \dots & \mathbf{a}_S^T(n) \end{bmatrix}^T$ of the dictionary are estimated using the penalized LS estimator, i.e.,

$$\hat{\mathbf{a}}(n) = \arg\min_{\mathbf{a}} \frac{1}{2} \left\| \mathbf{y}_n - \mathbf{W}\mathbf{a}(n) \right\|_2^2 + \lambda_{\mathrm{L}} \left\| \mathbf{a}(n) \right\|_1 + \lambda_{\mathrm{GL}} \sum_{r=1}^{S} \left\| \mathbf{a}_r(n) \right\|_2, \qquad (51)$$

where λ_{L} and λ_{GL} are the regularization coefficients of the penalties. Second, Q fundamental frequencies are estimated for a given $\tilde{Q}$ sources such as

$$\left\{ \hat{\omega}_{0,1} \quad \hat{\omega}_{0,2} \quad \dots \quad \hat{\omega}_{0,\tilde{Q}} \right\}_{\mathrm{PEBS}} = \arg \max_{\omega_{0,1}\, \omega_{0,2} \cdots \omega_{0,\tilde{Q}}} P\left(\{ \|\hat{\mathbf{a}}_r(n)\|_2 \, | \omega_{0,r} \}_{r=1}^{S} \right). \qquad (52)$$

The resulting estimates may suffer from some spurious estimates [1]. In paper C, we prove that the regularization coefficients should not be identical for all components of the dictionary, and we apply flexible penalty terms for a smooth spectral evaluation over multiple frames.

3.3 Performance Measures

The performance of the parameter estimates $\hat{\boldsymbol{\eta}}$ is tested in different situations of signal and noise. This evaluation is conducted to find uncertainties from the true parameter values. The squared error $(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta})^2$ has more emphasis on large errors than the absolute error $|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}|$. Therefore, we analyze statistics of the squared error, i.e., $\mathrm{E}\{(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta})^2\}$, by repeating the estimation process of the same signal. This sampling is also known as Monte-Carlo simulation

that results in random samples $\hat{\boldsymbol{\eta}}_b$ for $b = 1, 2, \ldots, B$ to calculate the MSE of B estimates from random samples, i.e.,

$$\mathrm{MSE}(\hat{\boldsymbol{\eta}}) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\boldsymbol{\eta}}_b - \boldsymbol{\eta})^2. \tag{53}$$

In Gaussian noise, the MSE of an unbiased estimator[2], i.e., $\mathrm{E}\{\hat{\boldsymbol{\eta}}\} = \boldsymbol{\eta}$, is evaluated by the Cramér-Rao lower bound (CRLB) [66], and the estimator is called statistically efficient when the MSE attains the boundary. The CRLB of the ith parameter of the parameter vector $\boldsymbol{\eta}$ is defined as

$$\mathrm{CRLB}([\hat{\boldsymbol{\eta}}]_i) = [\mathbf{I}(\boldsymbol{\eta})^{-1}]_{i,i}, \tag{54}$$

with

$$[\mathbf{I}(\boldsymbol{\eta})]_{i,j} = -\mathrm{E}\left\{\frac{\partial^2 \ln P\left(\left\{\left\{y_m(n)\right\}_{n=0}^{N-1}\right\}_{m=0}^{M-1}, \boldsymbol{\eta}\right)}{\partial[\boldsymbol{\eta}]_i\,\partial[\boldsymbol{\eta}]_j}\right\}, \tag{55}$$

where $\ln P\left(\left\{\left\{y_m(n)\right\}_{n=0}^{N-1}\right\}_{m=0}^{M-1}, \boldsymbol{\eta}\right)$ is the log-likelihood function of $M \times N$ independent samples. In white Gaussian noise, asymptotic CRLBs of joint fundamental frequency and DOA estimates have been formulated for ULAs in [61]. In paper B, we derive the CRLBs in colored Gaussian noise which are given by

$$\mathrm{CRLB}(\omega_0) = \frac{12}{NM(N^2-1)}(\mathbf{d}_L^T\mathbf{R}_{\Delta\boldsymbol{\Psi}}^{-1}\mathbf{d}_L)^{-1}, \tag{56}$$

$$\mathrm{CRLB}(\theta_0) = \Big[\frac{12c^2}{NM(M^2-1)(\omega_0 f_s \delta\cos\theta_0)^2} + \frac{12\tan^2\theta_0}{NM(N^2-1)\omega_0^2}\Big](\mathbf{d}_L^T\mathbf{R}_{\Delta\boldsymbol{\Psi}}^{-1}\mathbf{d}_L)^{-1}. \tag{57}$$

We also show that the CRLB of the DOA for the given fundamental frequency is lower than or equal to the CRLB of the DOA in the joint estimates, i.e., $\mathrm{CRLB}(\theta_0|\omega_0) \leq \mathrm{CRLB}(\theta_0)$, and the CRLB of the fundamental frequency for the given DOA is the same as the CRLB of the DOA in the joint estimates, i.e., $\mathrm{CRLB}(\omega_0|\theta_0) = \mathrm{CRLB}(\omega_0)$.

4 Contributions

This section gives an overview of the papers A through G which form the main contribution of this thesis. In multichannel noise reduction, paper A

[2]Note that the expected value of the squared error includes both variance and squared bias-error, i.e., $\mathrm{E}\{(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta})^2\} = \mathrm{Variance}(\hat{\boldsymbol{\eta}}) + \big[\mathrm{Bias}(\hat{\boldsymbol{\eta}})\big]^2$.

proposes a class of beamformers. The beamformers are designed using the fundamental frequency and the direction of arrival (DOA) of periodic signals. The estimation of the fundamental frequency and the DOA is considered in papers B, C, D, E, and F. Besides the proposed harmonic model-based beamformers, paper G also proposes a non-parametric beamformer with controllable linear constrains on the DOAs of the signal of interest and interferers.

Paper A: Harmonic model-based beamforming for speech enhancement
Beamformers should ideally be designed so that the desired signal, from a certain direction of arrival, is passed while background noise and interferers, at other directions, are attenuated as much as possible. Traditional broadband beamformers are designed without a prior assumption on the signal. In this paper, we propose a new class of broadband beamforming to pass the harmonics of periodic signals. The proposed harmonic model-based beamforming has advantages over the traditional broadband beamforming in noise reduction. As a result, the quality and intelligibility of speech signals are comparably higher than the results of traditional beamformers.

Paper B: Computationally efficient and noise robust fundamental frequency and DOA estimation
This paper proposes the estimation methods associated with noise statistics subject to distortionless constraints on the frequencies of the harmonics of periodic signals. The estimators are based on the maximum-likelihood (ML) of the frequency and the corresponding DOA estimates of the harmonics that make them statistically efficient. The proposed estimators are robust against different types of noise that makes them applicable in real-life scenarios. In addition, the methods are computationally simpler than the nonlinear least squares estimator [61].

Paper C: Multi-pitch estimation and tracking
Multi-pitch estimation, without posing a detailed a priori assumption of periodic sources, is a challenging problem. In this paper, we apply a general dictionary consisting of feasible fundamental frequencies and the corresponding harmonics. By doing this, we incorporate a Bayesian prior and assign data-dependent regularization coefficients in the (multi-) pitch estimation using block sparsity (PEBS) approach [1]. This version of the PEBS method has advantages on spectral estimation with less bias error and no spurious pitch estimates.

Paper D: Pitch estimation and tracking
In paper B, we propose an estimator of the fundamental frequency from unconstrained frequency estimates (UFEs) of the harmonics. In a series of the ML estimates of a changing fundamental frequency, the frequency estimates must change smoothly over time. In this paper, we

propose two estimators, namely a hidden Markov model and a Kalman filter, to optimally use the correlation of the consecutive UFEs over time. The results show that the proposed Bayesian based estimators are more accurate and smoother than the result of the ML estimator.

Paper E: Joint fundamental frequency and DOA estimation using a broadband beamformer

Estimation of the DOA and the fundamental frequency is not an easy task in a scenario of multiple sources. In microphone array signal processing, beamforming is commonly applied to extract the signal of interest at the given DOA and frequency. In this paper, we propose a method to estimate the fundamental frequency and the DOA jointly from the output of the broadband minimum variance distortionless response (MVDR) beamformer [16]. This approach is faster than the other estimators. Moreover, the model order of the harmonics is also estimated from the output of the beamformer using the maximum a posteriori (MAP) model order estimation method.

Paper F: Joint fundamental frequency and model order estimation using a broadband beamformer

Fundamental frequency estimation methods are often based on a priori assumption on the model order of the signal. However, estimation of the model order is not trivial in scenarios with multiple sources. This paper proposes an estimator of the joint fundamental frequency and the model order from the output of a beamformer at the given DOA of the signal of interest. The estimator is based on the MAP model order estimation method in [102].

Paper G: A controllable linearly constrained beamformer

The linearly constrained minimum variance (LCMV) beamformer [46] is the known solution to reject interferers at the DOAs other than the signal of interest. This paper presents a general form of an optimal beamformer to compromise between noise reduction and interference rejection. To control the performance of the beamformer, we select some interferers either to reject or to attenuate using controllable constraints on the DOA of the audio sources. As a result, the proposed controllable LCMV (C-LCMV) beamformer has a performance between the MVDR and the LCMV beamformers.

5 Conclusion and Future Direction

In this thesis, we have contributed generally in noise reduction of periodic signals. The results of this research can be used to enhanace signals of

voiced speech and some musical instruments in different applications such as hearing-aids and music information retrival. We have exploited the model of periodic signals, and proposed some optimal solutions to estimate the signal of interest through estimating the parameters of the model. We have mainly focused on two steps. First, we have estimated the parameters of the signal of interest from noisy spatiotemporal samples captured by a microphone array. In array processing, a spatial filter is usually designed to steer a beam to the direction-of-arrival (DOA) of the signal of interest. Second, we have enhanced the signal of interest by introducing a framework to design spatiotemporal filters by exploiting the estimated parameters. The idea was tailored to extend a number of data-independent (fixed) and data-dependent (adaptive) beamformers using the estimated DOA and fundamental frequency of periodic signals. We have shown that the proposed model-based beamformers have advantages over the traditional non-parametric beamformers in increasing the quality and ineligibility of speech signals.

Fundamental frequency is the parameter of interest according to the model of harmonic signals. We have proposed a number of solutions to estimate the fundamental frequency and the DOA either separately or jointly. We have shown that the harmonic model-based beamformers are also capable of estimating the parameters. We have also proposed a joint estimator of the fundamental frequency and DOA from the output of a non-parametric beamformer. Moreover, we have proposed estimators of the parameters concerning the noise statistics based on the weighted least-squares (WLS) method. We have shown that this approach is a maximum likelihood estimator, which is statistically efficient in colored noise. The proposed WLS-based methods are computationally simpler than the state-of-the-art nonlinear least squares (NLS) estimator. With regard to the continuity of the frequency changes over time, we have also proposed two Bayesian methods in the fundamental frequency estimation. Although the aforementioned estimators make a priori assumptions on the number of harmonic sources and the model order of the harmonics, we have extended an estimator using a dictionary, which avoids such assumptions. We have shown that the estimates change over time smoothly.

Although many efforts have been dedicated to estimating the parameters of harmonic signals for several decades, parameter estimation of voiced speech signals is a difficult problem in the acoustic resonances of the vocal tract which are known as formants [76]. This phenomenon changes the spectral envelope of the spectrum in a smooth shape of peaks at formant frequencies and spectral valleys in other frequencies. This phenomenon may attenuate the power spectrum of some harmonics, which causes losses of some harmonics in low local SNRs. Therefore, estimation of the model order and the fundamental frequency of voiced speech signals would not be easy. The future work might investigate how to estimate the parameters of

harmonics with regard to the formant frequencies.

References

[1] S. I. Adalbjörnsson, A. Jakobsson, and M. G. Christensen, "Multi-pitch estimation exploiting block sparsity," *Signal Processing*, vol. 109, pp. 236–247, 2015.

[2] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Autom. Control*, vol. 19, no. 6, pp. 716–723, 1974.

[3] S. Applebaum and D. Chapman, "Adaptive arrays with main beam constraints," *IEEE Trans. Antennas Propag.*, vol. 24, no. 5, pp. 650–662, 1976.

[4] J. Benesty and Y. Huang, "A single-channel noise reduction MVDR filter," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2011, pp. 273–276.

[5] J. Benesty, Y. Huang, and J. Chen, *Microphone Array Signal Processing*. Springer-Verlag, 2008, vol. 1.

[6] J. Benesty, S. Makino, and J. Chen, Eds., *Speech Enhancement*, ser. Signals and Communication Technology. Springer, 2005.

[7] F. Beritelli, S. Casale, and A. Cavallaero, "A robust voice activity detector for wireless communications using soft computing," *IEEE J. Sel. Areas Commun.*, vol. 16, no. 9, pp. 1818–1829, 1998.

[8] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 4, 1979, pp. 208–211.

[9] N. Bertin, R. Badeau, and E. Vincent, "Enforcing harmonicity and smoothness in bayesian non-negative matrix factorization applied to polyphonic music transcription," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 18, no. 3, pp. 538–549, 2010.

[10] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[11] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[12] M. Brandstein and D. Ward, Eds., *Microphone Arrays - Signal Processing Techniques and Applications*. Springer-Verlag, 2001.

[13] M. S. Brandstein and H. Silverman, "A practical methodology for speech source localization with microphone arrays," *Comput. Speech Language*, 1997.

[14] K. Buckley, "Broad-band beamforming and the generalized sidelobe canceller," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 5, pp. 1322–1323, Oct. 1986.

[15] D. Byrd and T. Crawford, "Problems of music information retrieval in the real world," *Information Processing & Management*, vol. 38, no. 2, pp. 249–272, 2002.

[16] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.

[17] Y. Chan, R. Hattin, and J. B. Plant, "The least squares estimation of time delay and its use in signal detection," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 26, no. 3, pp. 217–222, Jun 1978.

[18] J. Chen, J. Benesty, Y. Huang, and E. J. Diethorn, "Fundamentals of noise reduction," in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. Huang, Eds. Springer-Verlag, 2008, ch. 43, pp. 843–871.

[19] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction Wiener filter," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, no. 4, pp. 1218–1234, Jul. 2006.

[20] J. Chen, J. Benesty, and Y. Huang, "Robust time delay estimation exploiting redundancy among multiple microphones," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 549–557, 2003.

[21] ——, "Time delay estimation in room acoustic environments: an overview," *EURASIP J. on Applied Signal Process.*, vol. 2006, pp. 170–170, 2006.

[22] T. Chou, "Frequency-independent beamformer with low response error," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 5, May 1995, pp. 2995–2998.

[23] M. G. Christensen, J. L. Højvang, A. Jakobsson, and S. H. Jensen, "Joint fundamental frequency and order estimation using optimal filtering," *EURASIP J. on Applied Signal Process.*, vol. 2011, no. 1, pp. 1–18, Jun. 2011.

[24] M. G. Christensen and A. Jakobsson, "Multi-pitch estimation," *Synthesis Lectures on Speech and Audio Process.*, vol. 5, no. 1, pp. 1–160, 2009.

[25] ——, "Optimal filter designs for separating and enhancing periodic signals," *IEEE Trans. Signal Process.*, vol. 58, no. 12, pp. 5969–5983, Dec. 2010.

[26] M. G. Christensen, A. Jakobsson, and S. H. Jensen, "Fundamental frequency estimation using the shift-invariance property," in *Rec. Asilomar Conf. Signals, Systems, and Computers*, Nov. 2007, pp. 631–635.

[27] ——, "Joint high-resolution fundamental frequency and order estimation," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 5, pp. 1635–1644, Jul. 2007.

[28] M. G. Christensen, J. H. Jensen, A. Jakobsson, and S. H. Jensen, "On optimal filter designs for fundamental frequency estimation," *IEEE Signal Process. Lett.*, vol. 15, pp. 745–748, 2008.

[29] M. G. Christensen, P. Stoica, A. Jakobsson, and S. H. Jensen, "Multi-pitch estimation," *Elsevier Signal Process.*, vol. 88, no. 4, pp. 972–983, Apr. 2008.

[30] M. G. Christensen, P. Vera-Candeas, S. D. Somasundaram, and A. Jakobsson, "Robust subspace-based fundamental frequency estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2008, pp. 101–104.

[31] M. G. Christensen, "Accurate estimation of low fundamental frequencies from real-valued measurements," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 21, no. 10, pp. 2042–2056, 2013.

[32] M. Christensen and J. Jensen, "Pitch estimation for non-stationary speech," in *Rec. Asilomar Conf. Signals, Systems, and Computers*, Nov 2014, pp. 1400–1404.

[33] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Process. Lett.*, vol. 9, no. 1, pp. 12–15, Jan. 2002.

[34] H. Cox, R. Zeskind, and T. Kooij, "Practical supergain," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 3, pp. 393–398, 1986.

[35] H. Cox, R. Zeskind, and M. Owen, "Robust adaptive beamforming," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 35, no. 10, pp. 1365–1376, Oct. 1987.

[36] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays - Signal Processing Techniques and Applications*, M. S. Brandstein and D. B. Ward, Eds. Springer-Verlag, 2001, ch. 8, pp. 157–180.

[37] P. Djuric, "A model selection rule for sinusoids in white gaussian noise," *IEEE Trans. Signal Process.*, vol. 44, no. 7, pp. 1744–1751, Jul 1996.

[38] Y. Doweck, A. Amar, and I. Cohen, "Joint model order selection and parameter estimation of chirps with harmonic components," *IEEE Trans. Signal Process.*, vol. 63, no. 7, pp. 1765–1778, 2015.

[39] L. Du, T. Yardibi, J. Li, and P. Stoica, "Review of user parameter-free robust adaptive beamforming algorithms," *Digital Signal Processing*, vol. 19, no. 4, pp. 567–582, Jul. 2009.

[40] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proceedings of the IEEE*, vol. 80, no. 10, pp. 1526 –1555, oct 1992.

[41] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[42] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 251–266, Jul. 1995.

[43] Y. Ephraim, D. Malah, and B.-H. Juang, "On the application of hidden Markov models for enhancing noisy speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 12, pp. 1846–1856, 1989.

[44] M. Feder and E. Weinstein, "Parameter estimation of superimposed signals using the EM algorithm," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 4, pp. 477–489, Apr. 1988.

[45] N. H. Fletcher and T. D. Rossing, *The Physics of Musical Instruments*, 2nd ed. Springer Science+Business Media, Inc., 1998.

[46] O. L. Frost, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, vol. 60, no. 8, pp. 926–935, Aug. 1972.

[47] T. Fukuda, O. Ichikawa, and M. Nishimura, "Long-term spectro-temporal and static harmonic features for voice activity detection," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 4, no. 5, pp. 834–844, 2010.

[48] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Signal Process.*, vol. 49, no. 8, pp. 1614–1626, 2001.

[49] ——, "Iterative and sequential Kalman filter-based speech enhancement algorithms," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 6, no. 4, pp. 373–385, 1998.

[50] A. M. Goberman and M. Blomgren, "Fundamental frequency change during offset and onset of voicing in individuals with Parkinson disease," *Journal of Voice*, vol. 22, no. 2, pp. 178–191, 2008.

[51] S. Godsill and M. Davy, "Bayesian harmonic models for musical pitch estimation and analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, May 2002, pp. 1769–1772.

[52] S. Hahn, *Hilbert Transforms in Signal Processing*. Artech House, Inc., 1996.

[53] J. H. Hansen, M. Clements *et al.*, "Constrained iterative speech enhancement with application to speech recognition," *IEEE Trans. Signal Process.*, vol. 39, no. 4, pp. 795–805, 1991.

[54] D. Havelock, S. Kuwano, and M. Vorländer, *Handbook of signal processing in acoustics*. Springer Science & Business Media, 2008.

[55] W. Hess, *Pitch Determination of Speech Signals - Algorithms and Devices*. Springer-Verlag, 1983.

[56] H. G. Hirsch and C. Ehrlicher, "Noise estimation techniques for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, May 1995, pp. 153–156.

[57] Y. Hu and P. C. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 11, no. 4, pp. 334–341, 2003.

[58] K. Itoh and M. Mizushima, "Environmental noise reduction based on speech/non-speech identification for hearing aids," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1. IEEE, 1997, pp. 419–422.

[59] J. R. Jensen, J. Benesty, M. G. Christensen, and S. H. Jensen, "Enhancement of single-channel periodic signals in the time-domain," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 20, no. 7, pp. 1948–1963, Sep. 2012.

[60] J. R. Jensen, M. G. Christensen, and S. H. Jensen, "Joint spatio-temporal filtering methods for DOA and fundamental frequency estimation," *IEEE Trans. Signal Process.*, 2009, submitted.

[61] ——, "Nonlinear least squares methods for joint DOA and pitch estimation," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 21, no. 5, pp. 923 –933, May 2013.

[62] ——, "Statistically efficient methods for pitch and DOA estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 3900–3904.

[63] J. R. Jensen, M. G. Christensen, J. Benesty, and S. H. Jensen, "Joint spatio-temporal filtering methods for DOA and fundamental frequency estimation," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 23, no. 1, pp. 174–185, 2015.

[64] W. Jin, X. Liu, M. Scordilis, and L. Han, "Speech enhancement using harmonic emphasis and adaptive comb filtering," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 18, no. 2, pp. 356–368, Feb 2010.

[65] J. M. Kates, "Classification of background noises for hearing-aid applications," *Journal of the Acoustical Society of America*, vol. 97, no. 1, pp. 461–470, 1995.

[66] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice Hall, Inc., 1993.

[67] F. Khalil, J. P. Jullien, and A. Gilloire, "Microphone array for sound pickup in teleconference systems," *J. Audio Eng. Soc.*, vol. 42, no. 9, pp. 691–700, 1994.

[68] A. Klapuri and M. Davy, *Signal Processing Methods for Music Transcription*. Springer Science+Business Media LLC, 2006.

[69] D. H. Klatt, "Review of text-to-speech conversion for english," *J. Acoust. Soc. Am.*, vol. 82, no. 3, pp. 737–793, 1987.

[70] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[71] H. Li, P. Stoica, and J. Li, "Computationally efficient parameter estimation for harmonic sinusoidal signals," *Elsevier Signal Process.*, vol. 80(9), pp. 1937–1944, 2000.

[72] J. Li and P. Stoica, *Robust adaptive beamforming*. John Wiley & Sons, 2005, vol. 88.

[73] J. S. Lim and A. V. Oppenheim, "All-pole modeling of degraded speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 26, no. 3, pp. 197–210, Jun. 1978.

[74] W. Liu and S. Weiss, "Design of frequency invariant beamformers for broadband arrays," *IEEE Trans. Signal Process.*, vol. 56, no. 2, pp. 855–860, 2008.

[75] J. A. Logemann, H. B. Fisher, B. Boshes, and E. R. Blonsky, "Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of Parkinson patients," *J. Speech and Hearing Disorders*, vol. 43, no. 1, pp. 47–57, 1978.

[76] P. Loizou, *Speech Enhancement: Theory and Practice*. CRC Press, 2007.

[77] F.-L. Luo, J. Yang, C. Pavlovic, and A. Nehorai, "Adaptive null-forming scheme in digital hearing aids," *IEEE Trans. Signal Process.*, vol. 50, no. 7, pp. 1583–1590, 2002.

[78] K. M. Malladi and R. V. Rajakumar, "Estimation of time-varying ar models of speech through Gauss-Markov modeling," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 6. IEEE, 2003, pp. VI–305.

[79] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

[80] R. J. McAulay and T. F. Quatieri, "Sinusoidal coding," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Elsevier Science B.V., 1995, ch. 4, pp. 121–173.

[81] ——, "Mid-rate coding based on a sinusoidal representation of speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 10. IEEE, 1985, pp. 945–948.

[82] J. H. McClellan, "Multidimensional spectral estimation," *Proc. IEEE*, vol. 70, no. 9, pp. 1029–1039, 1982.

[83] D. P. Morgan, E. B. George, L. T. Lee, and S. M. Kay, "Cochannel speaker separation by harmonic enhancement and suppression," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 5, pp. 407–424, 1997.

[84] A. Nehorai and B. Porat, "Adaptive comb filtering for harmonic signal enhancement," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 5, pp. 1124–1138, Oct. 1986.

[85] M. Noll, "Pitch determination of human speech by harmonic product spectrum, the harmonic sum, and a maximum likelihood estimate," in *Proc. Symp. Comput. Process. Commun.*, 1969, pp. 779–797.

[86] K. K. Paliwal and A. Basu, "A speech enhancement method based on Kalman filtering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 12, Apr. 1987, pp. 177–180.

[87] T. W. Parsons, "Separation of speech from interfering speech by means of harmonic selection," *J. Acoust. Soc. Am.*, vol. 60, no. 4, pp. 911–918, 1976.

[88] G. E. Peterson and H. L. Barney, "Control methods used in a study of the vowels," *J. Acoust. Soc. Am.*, vol. 24, no. 2, pp. 175–184, 1952.

[89] L. R. Rabiner, "On the use of autocorrelation analysis for pitch detection," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 25, no. 1, pp. 24–33, Feb. 1977.

[90] J. Ramirez, J. M. Górriz, and J. C. Segura, *Voice activity detection. fundamentals and speech recognition system robustness*. INTECH Open Access Publisher, 2007.

[91] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, no. 5, pp. 465–471, 1978.

[92] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, 2001, pp. 749–752 vol.2.

[93] R. Roy and T. Kailath, "ESPRIT - estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 7, pp. 984–995, Jul. 1989.

[94] H. Sameti, H. Sheikhzadeh, L. Deng, and R. L. Brennan, "Hmm-based strategies for enhancement of speech signals embedded in nonstationary noise," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 6, no. 5, pp. 445–455, 1998.

[95] A. Sangwan, M. Chiranth, H. Jamadagni, R. Sah, R. V. Prasad, and V. Gaurav, "Vad techniques for real-time speech transmission on the Internet," in *IEEE Int. Conf. High Speed Networks and Multimedia Communications*, 2002, pp. 46–50.

[96] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, Mar. 1986.

[97] M. R. Schroeder, "Apparatus for suppressing noise and distortion in communication signals," US Patent 3,180,936, Apr. 27, 1965.

[98] ——, "Processing of communications signals to reduce effects of noise," US Patent 3,403,224, Sep. 24, 1968.

[99] B. L. Sim, Y. C. Tong, J. S. Chang, and C. T. Tan, "A parametric formulation of the generalized spectral subtraction method," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 4, pp. 328–337, Jul. 1998.

[100] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. IEEE Workshop Appl. of Signal Process. to Aud. and Acoust.* IEEE, 2003, pp. 177–180.

[101] P. Stoica and R. Moses, *Spectral Analysis of Signals*. Pearson Education, Inc., 2005.

[102] P. Stoica and Y. Selen, "Model-order selection: a review of information criterion rules," *IEEE Signal Process. Mag.*, vol. 21, no. 4, pp. 36–47, Jul. 2004.

[103] J. Sward, J. Brynolfsson, A. Jakobsson, and M. Hansson-Sandsten, "Sparse semi-parametric estimation of harmonic chirp signals," *IEEE Trans. Signal Process.*, vol. PP, no. 99, pp. 1798–1807, 2015.

[104] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2010, pp. 4214–4217.

[105] J. Tabrikian, S. Dubnov, and Y. Dickalov, "Maximum a-posteriori probability pitch tracking in noisy environments using harmonic model," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 1, pp. 76 – 87, Jan. 2004.

[106] S. Tretter, "Estimating the frequency of a noisy sinusoid by linear regression (corresp.)," *IEEE Trans. Inf. Theory*, vol. 31, no. 6, pp. 832–835, 1985.

[107] H. L. Van Trees, *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*. John Wiley & Sons, Inc., 2002.

[108] B. D. Van Veen and K. M. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE ASSP Mag.*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[109] D. B. Ward, Z. Ding, R. Kennedy *et al.*, "Broadband doa estimation using frequency invariant beamforming," *IEEE Trans. Signal Process.*, vol. 46, no. 5, pp. 1463–1469, 1998.

[110] D. B. Ward, R. A. Kennedy, and R. C. Williamson, "Theory and design of broadband sensor arrays with frequency invariant far-field beam patterns," *J. Acoust. Soc. Am.*, vol. 97, no. 2, pp. 1023–1034, 1995.

[111] M. Wax and T. Kailath, "Optimum localization of multiple sources by passive arrays," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 31, no. 5, pp. 1210–1217, 1983.

[112] D. P. Welker, J. E. Greenberg, J. G. Desloge, and P. M. Zurek, "Microphone-array hearing aids with binaural output. ii. a two-microphone adaptive system," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 6, pp. 543–551, 1997.

[113] M. Wohlmayr and M. Képesi, "Joint position-pitch extraction from multichannel audio," in *Proc. Interspeech*, Aug. 2007, pp. 1629–1632.

[114] T. Yardibi, J. Li, P. Stoica, M. Xue, and A. B. Baggeroer, "Source localization and sensing: A nonparametric iterative adaptive approach based on weighted least squares," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 46, no. 1, pp. 425–443, Jan. 2010.

[115] J. X. Zhang, M. G. Christensen, S. H. Jensen, and M. Moonen, "Joint DOA and multi-pitch estimation based on subspace techniques," *EURASIP J. on Advances in Signal Processing*, vol. 2012, no. 1, pp. 1–11, Jan. 2012.

Part II

Papers

Paper A

A Class of Parametric Broadband Beamformers Based on the Fundamental Frequency

Sam Karimian-Azari, Jesper Rindom Jensen, Jacob Benesty,
and Mads Græsbøll Christensen

The paper has been submitted in the
The Journal of the Acoustical Society of America, 2016.

In peer-review
The layout has been revised.

Abstract

Broadband beamforming is a well-known solution to multichannel noise reduction. In noise reduction, speech quality is directly related to the amount of residual noise and speech distortion. This paper presents a framework for parametric broadband beamforming which exploits the frequency-domain sparsity of voiced speech to achieve more noise reduction than traditional nonparametric broadband beamfoming without introducing additional distortion. In this approach, the harmonic modeling of voiced speech signals is considered to parameterize the beamformers specifically by the fundamental frequency of the harmonics. This approach considers separation and enhancement of periodic sources by exploiting the spectral and spatial properties of the signal sources. Accordingly, both data-independent and data-dependent harmonic model-based beamformers are derived in the time domain, i.e., (1) delay-and-sum, (2) null forming, (3) Wiener, (4) minimum variance distortionless response (MVDR), and (5) linearly constrained minimum variance beamformers. In addition, this paper also introduces a spatiotemporal filter as a trade-off between the maximum signal-to-noise ratio solution and the proposed harmonic model-based MVDR beamformer. Some numerical results on synthetic signals and real-life examples confirm the superior properties of the introduced framework, in terms of noise reduction, speech distortion, and objective measures for speech quality and speech intelligibility, compared to nonparametric broadband beamformers.

1 Introduction

Speech signals recorded by voice communication systems are often accompanied by unwanted noise and interferences in real life. These nuisance signals, that degrade the quality and intelligibility of speech signals, have a profound impact on voice communication systems, so an effective speech enhancement method is required to mitigate or eliminate the effects of added noise and interference. Nowadays, many voice communication systems are equipped with microphone arrays that provide spatial sampling in addition to the temporal sampling. Microphone arrays increases the performance of voice communication systems as the number of microphones increases, since the noise reduction and the degrees of freedom to separate interferers are potentially increased [1].

Beamforming is one approach to noise reduction using microphone arrays. It comprises of a set of finite-impulse-response (FIR) filters to create a space-tapered or a spatiotemporal filter, and an optimal filter is desired to minimize the noise and competing interference with a reasonable distortion on the desired speech signal, which can be obtained with, e.g., the multichannel Wiener filter. A beamformer is applied on multichannel signals to discriminate against signals from different direction of arrivals (DOAs), other

than that of the desired signal [2]. Narrowband beamformers, which are generally applied for communication and radar signals at a certain frequency band, attenuate signals from other directions. They are designed to pass the signal of interest and reject interferers [3]. Broadband beamformers are generally designed using narrowband beamformers for each of the frequency bands of the signal. To accomplish noncoherent noise reduction as well, numerous data-dependent beamformers have been developed (see [1, 4–6] and the references therein) which have been inspired mostly from single-channel data-dependent filters based on statistics of the signal and noise. The linearly constrained minimum variance (LCMV) beamformer [3] minimizes the residual noise, and enforces a set of linear constraints on the desired signal and interferers. Also, the Wiener post-filtering of the output of the minimum variance distortionless response (MVDR) beamformer [7] provides a minimum mean-squared error (MMSE) solution [8]. In general, nonparametric broadband beamformers are designed at all frequency bands. However, large parts of audio and speech signals are relatively sparse over the frequency bands, e.g., the harmonics of voiced speech. In other words, only a few frequency bands constitute the signal, and nonparametric broadband beamformers, e.g., the delay-and-sum and the MVDR beamformers, may partially retain noise in the frequency bands where the signal is zero.

For voiced speech and some musical instruments, it is reasonable to assume periodicity in short time intervals. Hence, the harmonic model, as the sum of sinusoids which are represented by a fundamental frequency and frequencies of the corresponding harmonics, can provide an efficient solution to capture, code, and transmit as well as manipulate and enhance periodic signals. Various harmonic model-based filters have been proposed for single-channel signal enhancement [9] and dereverberation [10]. For example, the data-dependent filter based on the optimal Capon spectral estimator [11] has been proposed with the distortionless constraint on the harmonics [12]. This harmonic model-based filtering passes the periodic signal of interest undistorted, and minimizes the noise and the other remaining interferers. However, it has still not been thoroughly considered how to optimally use the harmonic model for the enhancement of multichannel signals.

In the following, we exploit multichannel signals in order to increase degrees of freedom of the harmonic model-based filters, and suppress interferers which may be at the same frequency bands of the signal of interest in a scenario that interferers located at different positions. In this paper, we introduce optimum solutions to the multichannel signal enhancement in the maximum likelihood sense to provide the best possible output signal-to-noise ratio (SNR) for broadband periodic signals such as voiced speech. We propose harmonic model-based beamformers, in contrast to the nonparametric broadband beamformers. More specifically, we generalize the principles of the single-channel filterbank [12] and the spatiotemporal filtering

technique [13], and propose harmonic model-based beamforming which resembles a filterbank designed for the given spatial and spectral information. In this paper, the DOA and the fundamental frequency with the corresponding harmonics are treated as known parameters. The estimation problem of those parameters from noisy observed signals is outside the scope of this paper, but interested readers can find some existing methods for obtaining those in [13–20] and the references therein. We design fixed, or data-independent, delay-and-sum and null forming beamformers herein with the distortionless constraints on the aforementioned spatial and spectral parameters of the multichannel signals. To reduce noncoherent noise as well as the coherent interferers, we derive data-dependent harmonic model-based beamformers based on the nonparametric MVDR and LCMV beamformers and the multichannel Wiener filter. Moreover, the Karhunen-Loève expansion (KLE) is interesting as another data-dependent approach in noise reduction [21]. The KLE is computed in the subspace from the eigenvalue decomposition of the signal correlation matrix. The multichannel linear filtering technique has been associated with the KLE approaches based on the joint diagonalization [22] of either the correlation matrices of the noisy speech and the noise signals [23], or the correlation matrices of the speech and the noise signals [24]. The filters have been designed to minimize the speech distortion subject to a flexible noise reduction level [25], which results a trade-off distortion of the desired signal [23, 26]. In this paper, we also propose a linear filter based on the joint diagonalization of the correlation matrices of the speech and the noise signals. We apply the correlation matrix of the speech signals derived from properties of the harmonic signals instead of an estimate of the signals' correlation matrix. The amount of noise reduction and speech distortion depends on the number of applied eigenvectors in the proposed trade-off filter which compromises between the maximum SNR solution and the proposed harmonic model-based MVDR beamformer.

The remainder of this paper is organized as follows. Section II describes the multichannel signal model and problem formulation which form the basis of the paper. Section III outlines the conventional way of beamforming. Section IV represents the objective performance metrics of beamformers, namely the noise reduction factor, speech distortion index, and mean-squared error criterion. Then, Sections V and VI develop fixed and data-dependent harmonic model-based beamformers respectively. Section VII represents traditional non-parametric beamformers as a special case of the harmonic model-based beamformers. Then, some numerical examples are presented in Section VIII. Finally, Section IX concludes this work.

2 Signal Model and Problem Formulation

We consider the conventional signal model in which a microphone array with M sensors receives the unknown speech source signal $s(t)$, at the discrete-time index t, in some noise field. The received signals are expressed as [1]

$$\begin{aligned} y_m(t) &= g_m(t) * s(t) + v_m(t) \\ &= x_m(t) + v_m(t),\ m = 1, 2, \ldots, M, \end{aligned} \tag{A.1}$$

where $g_m(t)$ is the acoustic impulse response from the speech signal source to the mth microphone, $*$ denotes the convolution operation, and $x_m(t)$ and $v_m(t)$ are the speech and additive noise signals, respectively, at microphone m, which are assumed to be uncorrelated, and zero-mean. By definition, the terms $x_m(t),\ m = 1, 2, \ldots, M$, are coherent across the array. The noise signals, $v_m(t),\ m = 1, 2, \ldots, M$, are typically only partially coherent across the array. We further assume that microphone 1 is chosen as the reference sensor. Therefore, $x_1(t)$ is the desired signal that we want to recover from the sensors' observations from the far-field speaker. Moreover, we assume that the unknown speech source signal is quasi-stationary over a short interval, e.g., 20–30 ms. Hence, over the most recent time samples, $[\, s(t)\ \ s(t-1)\ \cdots\ s(t-L+1)\,]$, the spectral and statistical properties of the signal are constant for small L.

In this study, we consider a uniform linear array (ULA) consisting of M omnidirectional microphones, where the distance between two successive sensors is equal to δ and the direction of the source signal to this ULA is parameterized by the azimuthal angle θ lying inside the range 0 to π. The speech signal can be modeled by using the sum of sinusoids in the periods of voiced speech. Therefore, we model the convolved speech signal at the mth microphone as a harmonic signal source. Moreover, the acoustic impulse response essentially models the reverberation of an acoustic environment which leads to spectral and temporal smearing of the signal source. Although some nonharmonic components are added by the room reverberation due to indirect-path responses and long-term non-stationarity, the reverberation does not actually impair the frequency of the direct-path signal [10]. Therefore, by reconstructing the harmonic components and suppressing the residual noise and nonharmonic components we can enhance the signal of interest without assuming a priori knowledge about the indirect-path responses of the acoustic impulse response. For notational simplicity and computational efficiency, we use the discrete-time analytical signal [27] in an anechoic acoustic environment as well as in [20]:

$$x_m(t) = \sum_{n=1}^{N} a_n e^{\jmath n \omega_0 [t - f_s \tau_m(\theta)]}, \tag{A.2}$$

where N is the model order, the complex amplitude a_n is associated with the nth harmonic, $\jmath = \sqrt{-1}$ is the imaginary unit, ω_0 is the pitch or fundamental frequency, f_s is the sampling frequency,

$$\tau_m(\theta) = (m-1)\frac{\delta \cos\theta}{c} \tag{A.3}$$

is the relative delay of an impinging plane wave on the ULA, and c is the speed of sound in the air. Basically, the broadband signal, $x_m(t)$, whose fundamental frequency is ω_0, is the sum of N narrowband signals. Using (A.2), we can express (A.1) as

$$\begin{aligned} y_m(t) &= \sum_{n=1}^{N} a_n e^{\jmath n\omega_0[t - f_s\tau_m(\theta)]} + v_m(t) \\ &= \sum_{n=1}^{N} a_n e^{\jmath n\omega_0 t} e^{-\jmath n\omega_0 f_s \tau_m(\theta)} + v_m(t). \end{aligned} \tag{A.4}$$

Putting together the samples of the mth microphone observations in a vector of length L, we get

$$\begin{aligned} \mathbf{y}_m(t) &= \begin{bmatrix} y_m(t) & y_m(t-1) & \cdots & y_m(t-L+1) \end{bmatrix}^T \\ &= \mathbf{x}_m(t) + \mathbf{v}_m(t) \\ &= \mathbf{D}_{m,N}(\theta, \omega_0)\mathbf{a}(t, \omega_0) + \mathbf{v}_m(t), \end{aligned} \tag{A.5}$$

where the superscript ^{T} is the transpose operator, $\mathbf{x}_m(t) = \mathbf{D}_{m,N}(\theta, \omega_0)\mathbf{a}(t, \omega_0)$,

$$\mathbf{D}_{m,N}(\theta, \omega_0) = \begin{bmatrix} \mathbf{d}_{m,1}(\theta, \omega_0) & \mathbf{d}_{m,2}(\theta, \omega_0) & \cdots & \mathbf{d}_{m,N}(\theta, \omega_0) \end{bmatrix} \tag{A.6}$$

is a matrix of size $L \times N$, with

$$\mathbf{d}_{m,n}(\theta, \omega_0) = e^{-\jmath n\omega_0 f_s \tau_m(\theta)} \times \begin{bmatrix} 1 & e^{-\jmath n\omega_0} & \cdots & e^{-\jmath n\omega_0 (L-1)} \end{bmatrix}^T \tag{A.7}$$

being a vector of length L,

$$\mathbf{a}(t, \omega_0) = \begin{bmatrix} a_1 e^{\jmath\omega_0 t} & a_2 e^{\jmath 2\omega_0 t} & \cdots & a_N e^{\jmath N\omega_0 t} \end{bmatrix}^T \tag{A.8}$$

is a vector of length N, and

$$\mathbf{v}_m(t) = \begin{bmatrix} v_m(t) & v_m(t-1) & \cdots & v_m(t-L+1) \end{bmatrix}^T. \tag{A.9}$$

The complex amplitudes, $\begin{bmatrix} a_1 & a_2 & \cdots & a_N \end{bmatrix}$, are assumed to be zero-mean circular complex random variables that have independent phases uniformly distributed on the interval $(-\pi, \pi]$. Therefore $E\left[a_i a_j^*\right] = 0$ for $i \neq j$, and the correlation matrix of $\mathbf{a}$ (of size $N \times N$) is

$$\mathbf{R}_{\mathbf{a}} = \operatorname{diag}\left(E\left[|a_1|^2\right], E\left[|a_2|^2\right], \ldots, E\left[|a_N|^2\right]\right) \tag{A.10}$$

where $E[\cdot]$ is the mathematical expectation, and the superscript * is the complex-conjugate operator. Define the vector of length N:

$$\mathbf{1}_N = \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix}^T. \tag{A.11}$$

It is obvious that $\mathbf{1}_N^T \mathbf{a}(t, \omega_0) = x_1(t)$, which is the desired signal. Now, concatenating all microphone signal vectors, we obtain the vector of length ML:

$$\begin{aligned} \underline{\mathbf{y}}(t) &= \begin{bmatrix} \mathbf{y}_1^T(t) & \mathbf{y}_2^T(t) & \cdots & \mathbf{y}_M^T(t) \end{bmatrix}^T \\ &= \underline{\mathbf{x}}(t) + \underline{\mathbf{v}}(t) \\ &= \underline{\mathbf{D}}_N(\theta, \omega_0)\mathbf{a}(t, \omega_0) + \underline{\mathbf{v}}(t), \end{aligned} \tag{A.12}$$

where $\underline{\mathbf{x}}(t) = \underline{\mathbf{D}}_N(\theta, \omega_0)\mathbf{a}(t, \omega_0)$,

$$\underline{\mathbf{D}}_N(\theta, \omega_0) = \begin{bmatrix} \mathbf{D}_{1,N}(\theta, \omega_0) \\ \mathbf{D}_{2,N}(\theta, \omega_0) \\ \vdots \\ \mathbf{D}_{M,N}(\theta, \omega_0) \end{bmatrix} \tag{A.13}$$

is a matrix of size $ML \times N$, and

$$\underline{\mathbf{v}}(t) = \begin{bmatrix} \mathbf{v}_1^T(t) & \mathbf{v}_2^T(t) & \cdots & \mathbf{v}_M^T(t) \end{bmatrix}^T. \tag{A.14}$$

We deduce that the correlation matrix of $\underline{\mathbf{y}}(k)$ (of size $ML \times ML$) is

$$\begin{aligned} \mathbf{R}_{\underline{\mathbf{y}}} &= E\left[\underline{\mathbf{y}}(t)\underline{\mathbf{y}}^H(t)\right] \\ &= \mathbf{R}_{\underline{\mathbf{x}}} + \mathbf{R}_{\underline{\mathbf{v}}} \\ &= \underline{\mathbf{D}}_N(\theta, \omega_0)\mathbf{R}_{\mathbf{a}}\underline{\mathbf{D}}_N^H(\theta, \omega_0) + \mathbf{R}_{\underline{\mathbf{v}}}, \end{aligned} \tag{A.15}$$

where the superscript ^{H} is the conjugate-transpose operator,
$\mathbf{R}_{\underline{\mathbf{x}}} = \underline{\mathbf{D}}_N(\theta, \omega_0)\mathbf{R}_{\mathbf{a}}\underline{\mathbf{D}}_N^H(\theta, \omega_0)$ is the correlation matrix of $\underline{\mathbf{x}}(t)$, and $\mathbf{R}_{\underline{\mathbf{v}}} = E\left[\underline{\mathbf{v}}(t)\underline{\mathbf{v}}^H(t)\right]$ is the correlation matrix of $\underline{\mathbf{v}}(t)$. It is important to observe that the matrix $\mathbf{R}_{\underline{\mathbf{x}}}$ is rank deficient only if $ML > N$, which is easy to satisfy by just increasing M or (especially) L; this will always be assumed. We will see how to exploit the nullspace of $\mathbf{R}_{\underline{\mathbf{x}}}$ to derive all kind of broadband beamformers. In the rest, it is assumed that the desired signal propagates from the fixed direction θ_0; so in (A.12) and (A.15), θ is replaced by θ_0. Therefore, our signal model is now

$$\underline{\mathbf{y}}(t) = \underline{\mathbf{D}}_N(\theta_0, \omega_0)\mathbf{a}(t, \omega_0) + \underline{\mathbf{v}}(t). \tag{A.16}$$

3 Broadband Beamforming

The conventional way to perform beamforming is by applying a complex-valued temporal linear filter of length L at the output of each microphone and summing the filtered signals. The beamformer output is then

$$\begin{aligned} z(t) &= \sum_{m=1}^{M} \mathbf{h}_m^H \mathbf{y}_m(t) \\ &= \underline{\mathbf{h}}^H \underline{\mathbf{y}}(t) \\ &= x_{\mathrm{fd}}(t) + v_{\mathrm{rn}}(t), \end{aligned} \tag{A.17}$$

where

$$\underline{\mathbf{h}} = \left[\begin{array}{cccc} \mathbf{h}_1^T & \mathbf{h}_2^T & \cdots & \mathbf{h}_M^T \end{array} \right]^T \tag{A.18}$$

is the spatiotemporal linear filter of length ML, with $\mathbf{h}_m$, $m = 1, 2, \ldots, M$ being the temporal filters of length L,

$$\begin{aligned} x_{\mathrm{fd}}(t) &= \sum_{m=1}^{M} \mathbf{h}_m^H \mathbf{D}_{m,N}(\theta_0, \omega_0) \mathbf{a}(t, \omega_0) \\ &= \underline{\mathbf{h}}^H \underline{\mathbf{D}}_N(\theta_0, \omega_0) \mathbf{a}(t, \omega_0) \end{aligned} \tag{A.19}$$

is the filtered desired signal, and

$$\begin{aligned} v_{\mathrm{rn}}(t) &= \sum_{m=1}^{M} \mathbf{h}_m^H \mathbf{v}_m(t) \\ &= \underline{\mathbf{h}}^H \underline{\mathbf{v}}(t) \end{aligned} \tag{A.20}$$

is the residual noise. We deduce that the variance of $z(t)$ is

$$\begin{aligned} \sigma_z^2 &= \underline{\mathbf{h}}^H \mathbf{R}_{\underline{\mathbf{y}}} \underline{\mathbf{h}} \\ &= \sigma_{x_{\mathrm{fd}}}^2 + \sigma_{v_{\mathrm{rn}}}^2, \end{aligned} \tag{A.21}$$

where

$$\sigma_{x_{\mathrm{fd}}}^2 = \underline{\mathbf{h}}^H \underline{\mathbf{D}}_N(\theta_0, \omega_0) \mathbf{R}_{\mathbf{a}} \underline{\mathbf{D}}_N^H(\theta_0, \omega_0) \underline{\mathbf{h}} \tag{A.22}$$

is the variance of $x_{\mathrm{fd}}(t)$ and

$$\sigma_{v_{\mathrm{rn}}}^2 = \underline{\mathbf{h}}^H \mathbf{R}_{\underline{\mathbf{v}}} \underline{\mathbf{h}} \tag{A.23}$$

is the variance of $v_{\mathrm{rn}}(t)$.

4 Performance Measures

In this section, we derive some very useful performance measures that are needed not only for the derivation of different kind of beamformers but also for their evaluation. The performance measures are special cases of the well-known general expressions in [1, 28] by using the harmonic model. We parameterize the signal correlation matrix, and discuss the noise reduction performance, as well as the speech distortion performance, and the mean-squared error (MSE) criterion. We show how the MSE is naturally related to all second-order performance measures.

4.1 Noise Reduction

Since microphone 1 is the reference, the input signal-to-noise ratio (SNR) is computed from the first L components of $\underline{\mathbf{y}}(t)$ as defined in (A.16), i.e., $\mathbf{y}_1(t) = \mathbf{D}_{1,N}(\theta_0, \omega_0)\mathbf{a}(t, \omega_0) + \mathbf{v}_1(t)$. We easily find that

$$\begin{aligned} \text{iSNR} &= \frac{\text{tr}\left[\mathbf{D}_{1,N}(\theta_0, \omega_0)\mathbf{R}_{\mathbf{a}}\mathbf{D}_{1,N}^H(\theta_0, \omega_0)\right]}{\text{tr}\left(\mathbf{R}_{\mathbf{v}_1}\right)} \\ &= \frac{\mathbf{1}_N^T\mathbf{R}_{\mathbf{a}}\mathbf{1}_N}{\sigma_{v_1}^2}, \end{aligned} \tag{A.24}$$

where $\text{tr}(\cdot)$ denotes the trace of a square matrix, $\mathbf{R}_{\mathbf{v}_1}$ is the correlation matrix of $\mathbf{v}_1(t)$, and $\sigma_{v_1}^2$ is the variance of $v_1(t)$.

The output SNR is obtained from (A.21). It is given by

$$\begin{aligned} \text{oSNR}\left(\underline{\mathbf{h}}\right) &= \frac{\sigma_{x_{\text{fd}}}^2}{\sigma_{v_{\text{rn}}}^2} \\ &= \frac{\underline{\mathbf{h}}^H\underline{\mathbf{D}}_N(\theta_0, \omega_0)\mathbf{R}_{\mathbf{a}}\underline{\mathbf{D}}_N^H(\theta_0, \omega_0)\underline{\mathbf{h}}}{\sigma_{v_1}^2\underline{\mathbf{h}}^H\mathbf{\Gamma}_{\underline{\mathbf{v}}}\underline{\mathbf{h}}}, \end{aligned} \tag{A.25}$$

where $\mathbf{\Gamma}_{\underline{\mathbf{v}}} = \mathbf{R}_{\underline{\mathbf{v}}}/\sigma_{v_1}^2$ is the pseudo-correlation matrix of $\underline{\mathbf{v}}(t)$. We see from (A.25) that the gain in SNR is

$$\begin{aligned} \mathcal{G}\left(\underline{\mathbf{h}}\right) &= \frac{\text{oSNR}\left(\underline{\mathbf{h}}\right)}{\text{iSNR}} \\ &= \frac{1}{\mathbf{1}_N^T\mathbf{R}_{\mathbf{a}}\mathbf{1}_N} \times \frac{\underline{\mathbf{h}}^H\underline{\mathbf{D}}_N(\theta_0, \omega_0)\mathbf{R}_{\mathbf{a}}\underline{\mathbf{D}}_N^H(\theta_0, \omega_0)\underline{\mathbf{h}}}{\underline{\mathbf{h}}^H\mathbf{\Gamma}_{\underline{\mathbf{v}}}\underline{\mathbf{h}}}. \end{aligned} \tag{A.26}$$

The white noise gain (WNG), $\mathcal{W}\left(\underline{\mathbf{h}}\right)$, is obtained by taking $\mathbf{\Gamma}_{\underline{\mathbf{v}}} = \mathbf{I}_{ML}$, where $\mathbf{I}_{ML}$ is the $ML \times ML$ identity matrix.

The noise reduction factor quantifies the amount of noise being attenuated by the beamformer. This quantity is defined as the ratio of the power of the original noise over the power of the noise remaining after filtering, i.e.,

$$\begin{aligned}\xi_{\text{nr}}\left(\underline{\mathbf{h}}\right) &= \frac{\text{tr}\left(\mathbf{R}_{\mathbf{v}_1}\right)}{L\sigma_{v_{\text{rn}}}^2} \\ &= \frac{1}{\underline{\mathbf{h}}^H\boldsymbol{\Gamma}_{\underline{\mathbf{v}}}\underline{\mathbf{h}}}.\end{aligned} \tag{A.27}$$

For optimal filters, it is desired that $\xi_{\text{nr}}\left(\underline{\mathbf{h}}\right) \geq 1$.

4.2 Speech Distortion

The desired speech signal can be distorted by the beamformer. Therefore, the speech reduction factor is defined as

$$\begin{aligned}\xi_{\text{sr}}\left(\underline{\mathbf{h}}\right) &= \frac{\text{tr}\left(\mathbf{R}_{\mathbf{x}_1}\right)}{L\sigma_{x_{\text{fd}}}^2} \\ &= \frac{\mathbf{1}_N^T\mathbf{R}_{\mathbf{a}}\mathbf{1}_N}{\underline{\mathbf{h}}^H\underline{\mathbf{D}}_N(\theta_0,\omega_0)\mathbf{R}_{\mathbf{a}}\underline{\mathbf{D}}_N^H(\theta_0,\omega_0)\underline{\mathbf{h}}}.\end{aligned} \tag{A.28}$$

For optimal filters, it is preferred that $\xi_{\text{sr}}\left(\underline{\mathbf{h}}\right) \geq 1$. In the distortionless case, we have $\xi_{\text{sr}}\left(\underline{\mathbf{h}}\right) = 1$. Hence, a beamformer that does not affect the desired signal requires the constraint:

$$\underline{\mathbf{h}}^H\underline{\mathbf{D}}_N(\theta_0,\omega_0) = \mathbf{1}_N^T. \tag{A.29}$$

It is clear that we always have

$$\mathcal{G}\left(\underline{\mathbf{h}}\right) = \frac{\xi_{\text{nr}}\left(\underline{\mathbf{h}}\right)}{\xi_{\text{sr}}\left(\underline{\mathbf{h}}\right)}. \tag{A.30}$$

The distortion can also be measured with the speech distortion index:

$$\begin{aligned}v_{\text{sd}}\left(\underline{\mathbf{h}}\right) &= L\frac{E\left[\left|x_{\text{fd}}(t)-x_1(t)\right|^2\right]}{\text{tr}\left(\mathbf{R}_{\mathbf{x}_1}\right)} \\ &= \frac{E\left[\left|\underline{\mathbf{h}}^H\underline{\mathbf{D}}_N(\theta_0,\omega_0)\mathbf{a}(t,\omega_0)-\mathbf{1}_N^T\mathbf{a}(t,\omega_0)\right|^2\right]}{\mathbf{1}_N^T\mathbf{R}_{\mathbf{a}}\mathbf{1}_N}. \\ &= \frac{\left[\underline{\mathbf{h}}^H\underline{\mathbf{D}}_N(\theta_0,\omega_0)-\mathbf{1}_N^T\right]\mathbf{R}_{\mathbf{a}}\left[\underline{\mathbf{D}}_N^H(\theta_0,\omega_0)\underline{\mathbf{h}}-\mathbf{1}_N\right]}{\mathbf{1}_N^T\mathbf{R}_{\mathbf{a}}\mathbf{1}_N}.\end{aligned} \tag{A.31}$$

It has been proven in [29] that $0 \leq v_{\text{sd}}\left(\underline{\mathbf{h}}\right) \leq 1$, and a value of $v_{\text{sd}}\left(\underline{\mathbf{h}}\right)$ close to 0 is preferred for optimal filters.

4.3 Mean-Squared Error Criterion

We define the error signal between the estimated and desired signals as

$$\begin{aligned} e(t) &= z(t) - x_1(t) \\ &= e_{\text{ds}}(t) + e_{\text{rs}}(t), \end{aligned} \tag{A.32}$$

where

$$\begin{aligned} e_{\text{ds}}(t) &= x_{\text{fd}}(t) - x_1(t) \\ &= \left[\underline{\mathbf{h}}^H \underline{\mathbf{D}}_N(\theta_0, \omega_0) - \mathbf{1}_N^T\right] \mathbf{a}(t, \omega_0) \end{aligned} \tag{A.33}$$

represents the signal distortion and $e_{\text{rs}}(t) = v_{\text{rn}}(t)$ represents the residual noise. We deduce that the mean-squared error (MSE) criterion is

$$\begin{aligned} J(\underline{\mathbf{h}}) = &\ E\left[|e(t)|^2\right] \\ = &\ \mathbf{1}_N^T \mathbf{R}_{\mathbf{a}} \mathbf{1}_N + \underline{\mathbf{h}}^H \underline{\mathbf{D}}_N(\theta_0, \omega_0) \mathbf{R}_{\mathbf{a}} \underline{\mathbf{D}}_N^H(\theta_0, \omega_0) \underline{\mathbf{h}} \\ &- \underline{\mathbf{h}}^H \underline{\mathbf{D}}_N(\theta_0, \omega_0) \mathbf{R}_{\mathbf{a}} \mathbf{1}_N - \mathbf{1}_N^T \mathbf{R}_{\mathbf{a}} \underline{\mathbf{D}}_N^H(\theta_0, \omega_0) \underline{\mathbf{h}} + \underline{\mathbf{h}}^H \mathbf{R}_{\underline{\mathbf{v}}} \underline{\mathbf{h}}. \end{aligned} \tag{A.34}$$

Since $E\left[e_{\text{ds}}(t) e_{\text{rs}}^*(t)\right] = 0$, $J(\underline{\mathbf{h}})$ can also be expressed as

$$\begin{aligned} J(\underline{\mathbf{h}}) &= E\left[|e_{\text{ds}}(t)|^2\right] + E\left[|e_{\text{rs}}(t)|^2\right] \\ &= J_{\text{ds}}(\underline{\mathbf{h}}) + J_{\text{rs}}(\underline{\mathbf{h}}), \end{aligned} \tag{A.35}$$

where $J_{\text{ds}}(\underline{\mathbf{h}}) = \text{tr}(\mathbf{R}_{\mathbf{x}_1}) v_{\text{sd}}(\underline{\mathbf{h}}) / L$, and $J_{\text{rs}}(\underline{\mathbf{h}}) = \text{tr}(\mathbf{R}_{\mathbf{v}_1}) / L \xi_{\text{nr}}(\underline{\mathbf{h}})$. Finally, we have

$$\begin{aligned} \frac{J_{\text{ds}}(\underline{\mathbf{h}})}{J_{\text{rs}}(\underline{\mathbf{h}})} &= \text{iSNR} \times \xi_{\text{nr}}(\underline{\mathbf{h}}) \times v_{\text{sd}}(\underline{\mathbf{h}}) \\ &= \text{oSNR}(\underline{\mathbf{h}}) \times \xi_{\text{sr}}(\underline{\mathbf{h}}) \times v_{\text{sd}}(\underline{\mathbf{h}}). \end{aligned} \tag{A.36}$$

This shows how the MSEs are related to the most fundamental performance measures.

5 Fixed Harmonic Model-Based Beamformers

The harmonic model-based beamformers (HBs), which we introduce them throughout this paper, have distortionless constraints with respect to the harmonics, in contrast with the nonparametric broadband beamformers (BBs) which have distortionless constraints with respect to uniformly spaced frequency bands (we discuss about the BBs in Section 7). To understand the general idea, Fig. A.1 shows an example of the spatial directivity pattern of the beamformers for an equally spaced linear array. The vertical axis is

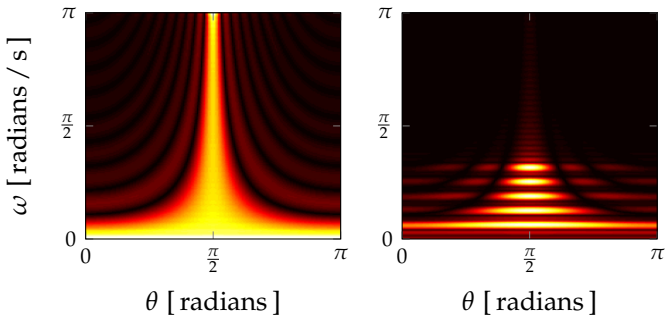


Fig. A.1: Spatial directivity pattern of two broadband beamformers designed with the constraints with respect to (left) all frequency bands and (right) harmonics.

the normalized frequency in radians per second, and the horizontal axis is the angle in radians. The beamformers have unit gain with respect to their constraints at the given DOA, i.e., $\theta_0 = \pi/2$, and the frequencies of the harmonics (right) and uniformly spaced frequency bands (left).

5.1 Delay-and-Sum

The delay-and-sum (DS) beamformer is obtained by maximizing the WNG subject to the distortionless constraint, i.e.,

$$\min_{\underline{\mathbf{h}}} \underline{\mathbf{h}}^H \underline{\mathbf{h}} \quad \text{subject to} \quad \underline{\mathbf{h}}^H \underline{\mathbf{D}}_N(\theta_0, \omega_0) = \mathbf{1}_N^T. \tag{A.37}$$

We deduce that the optimal solution is

$$\underline{\mathbf{h}}_{\text{DS-HB}} = \underline{\mathbf{D}}_N(\theta_0, \omega_0) \left[\underline{\mathbf{D}}_N^H(\theta_0, \omega_0) \underline{\mathbf{D}}_N(\theta_0, \omega_0) \right]^{-1} \mathbf{1}_N. \tag{A.38}$$

As a result, the WNG is

$$\mathcal{W}\left(\underline{\mathbf{h}}_{\text{DS-HB}}\right) = \frac{1}{\mathbf{1}_N^T \left[\underline{\mathbf{D}}_N^H(\theta_0, \omega_0) \underline{\mathbf{D}}_N(\theta_0, \omega_0) \right]^{-1} \mathbf{1}_N}. \tag{A.39}$$

In the presence of spatially white noise, the DS beamformer is optimal in the sense that it gives the maximum gain in SNR without distorting the desired signal. However, in the presence of other noises, we should not expect very high gains. Moreover, we can obtain $\lim_{ML \to \infty} \underline{\mathbf{D}}_N^H(\theta_0, \omega_0) \underline{\mathbf{D}}_N(\theta_0, \omega_0) = ML \times \mathbf{I}_N$. Therefore, the WNG of the DS-HB depends directly to both M and L, i.e., $\mathcal{W}\left(\underline{\mathbf{h}}_{\text{DS-HB}}\right) \to ML/N$.

5.2 Null Forming

Let us assume that there is a broadband interference with fundamental frequency ω_1 and model order N_1 in the direction θ_1. The matrix $\underline{\mathbf{D}}_{N_1}(\theta_1, \omega_1)$ of size $ML \times N_1$ is associated with this interference.

Now, we would like to perfectly recover the desired signal and completely cancel the interference. The constraint is then

$$\underline{\mathbf{h}}^H \underline{\mathbf{C}} = \left[\begin{array}{cc} \mathbf{1}_N^T & \mathbf{0}_{N_1}^T \end{array} \right], \tag{A.40}$$

where

$$\underline{\mathbf{C}} = \left[\begin{array}{cc} \underline{\mathbf{D}}_N(\theta_0, \omega_0) & \underline{\mathbf{D}}_{N_1}(\theta_1, \omega_1) \end{array} \right] \tag{A.41}$$

is the constraint matrix of size $ML \times (N + N_1)$ and $\mathbf{0}_{N_1}$ is the zero vector of length N_1. Then, our criterion is

$$\min_{\underline{\mathbf{h}}} \underline{\mathbf{h}}^H \underline{\mathbf{h}} \quad \text{subject to} \quad \underline{\mathbf{h}}^H \underline{\mathbf{C}} = \left[\begin{array}{cc} \mathbf{1}_N^T & \mathbf{0}_{N_1}^T \end{array} \right], \tag{A.42}$$

from which we find the optimal solution:

$$\underline{\mathbf{h}}_{\mathrm{NF-HB}} = \underline{\mathbf{C}} \left(\underline{\mathbf{C}}^H \underline{\mathbf{C}} \right)^{-1} \left[\begin{array}{c} \mathbf{1}_N \\ \mathbf{0}_{N_1} \end{array} \right]. \tag{A.43}$$

Obviously, we must have $ML > N + N_1$. The generalization of this approach to any number of interferences is straightforward.

6 Data-Dependent Harmonic Model-Based Beamformers

This section deals with a class of data-dependent beamformers, where some signal statistics need to be estimated. In theory, data-dependent beamformers give much better results than fixed beamformers since they can adjust pretty quickly to the new environment.

6.1 Wiener

The harmonic model-based Wiener beamformer is easily derived by taking the gradient of the MSE, $J(\underline{\mathbf{h}})$ [eq. (A.34)], with respect to $\underline{\mathbf{h}}$ and equating the result to zero:

$$\underline{\mathbf{h}}_{\mathrm{W-HB}} = \left[\underline{\mathbf{D}}_N(\theta_0, \omega_0) \mathbf{R}_{\mathbf{a}} \underline{\mathbf{D}}_N^H(\theta_0, \omega_0) + \mathbf{R}_{\underline{\mathbf{v}}} \right]^{-1} \underline{\mathbf{D}}_N(\theta_0, \omega_0) \mathbf{R}_{\mathbf{a}} \mathbf{1}_N. \tag{A.44}$$

Determining the matrix inverse with the Woodbury identity leads to another interesting formulation of the harmonic model-based Wiener beamformer:

$$\begin{aligned}\underline{\mathbf{h}}_{\mathrm{W-HB}} &= \mathbf{R}_{\underline{\mathbf{v}}}^{-1}\underline{\mathbf{D}}_N(\theta_0,\omega_0)\left[\mathbf{R}_{\mathbf{a}}^{-1}+\underline{\mathbf{D}}_N^H(\theta_0,\omega_0)\mathbf{R}_{\underline{\mathbf{v}}}^{-1}\underline{\mathbf{D}}_N(\theta_0,\omega_0)\right]^{-1}\mathbf{1}_N \\ &= \mathbf{R}_{\underline{\mathbf{v}}}^{-1}\underline{\mathbf{D}}_N(\theta_0,\omega_0)\left[\underline{\mathbf{D}}_N^H(\theta_0,\omega_0)\mathbf{R}_{\underline{\mathbf{v}}}^{-1}\underline{\mathbf{D}}_N(\theta_0,\omega_0)\right]^{-1}\mathbf{P}\left(\theta_0,\omega_0\right)\mathbf{1}_N,\end{aligned} \tag{A.45}$$

where

$$\mathbf{P}\left(\theta_0,\omega_0\right) = \left(\mathbf{R}_{\mathbf{a}}^{-1}\left[\underline{\mathbf{D}}_N^H(\theta_0,\omega_0)\mathbf{R}_{\underline{\mathbf{v}}}^{-1}\underline{\mathbf{D}}_N(\theta_0,\omega_0)\right]^{-1}+\mathbf{I}_N\right)^{-1}.$$

In spatially white noise, we can approximate $\mathbf{P}\left(\theta_0,\omega_0\right)$ as $\mathbf{P} = \left(\frac{\sigma_{v_1}^2}{ML}\mathbf{R}_{\mathbf{a}}^{-1}+\mathbf{I}_N\right)^{-1}$ for a large filter, i.e., $ML \to \infty$.

6.2 Minimum Variance Distortionless Response

The celebrated minimum variance distortionless response (MVDR) beamformer proposed by Capon [7, 11] is easily derived by optimizing the following criterion:

$$\min_{\underline{\mathbf{h}}} \underline{\mathbf{h}}^H\mathbf{R}_{\underline{\mathbf{v}}}\underline{\mathbf{h}} \quad \text{subject to} \quad \underline{\mathbf{h}}^H\underline{\mathbf{D}}_N(\theta_0,\omega_0) = \mathbf{1}_N^T. \tag{A.46}$$

We obtain

$$\underline{\mathbf{h}}_{\mathrm{MVDR-HB}} = \mathbf{R}_{\underline{\mathbf{v}}}^{-1}\underline{\mathbf{D}}_N(\theta_0,\omega_0)\left[\underline{\mathbf{D}}_N^H(\theta_0,\omega_0)\mathbf{R}_{\underline{\mathbf{v}}}^{-1}\underline{\mathbf{D}}_N(\theta_0,\omega_0)\right]^{-1}\mathbf{1}_N. \tag{A.47}$$

The perfectly matched beamformer to the signal parameters results in $\underline{\mathbf{h}}_{\mathrm{MVDR-HB}}^H\mathbf{R}_{\underline{\mathbf{y}}}\underline{\mathbf{h}}_{\mathrm{MVDR-HB}} = \mathbf{1}_N^T\mathbf{R}_{\mathbf{a}}\mathbf{1}_N + \underline{\mathbf{h}}_{\mathrm{MVDR-HB}}^H\mathbf{R}_{\underline{\mathbf{v}}}\underline{\mathbf{h}}_{\mathrm{MVDR-HB}}$. Therefore, minimizing the residual noise is equivalent to minimizing the noisy signal, i.e., $\underline{\mathbf{h}}^H\mathbf{R}_{\underline{\mathbf{y}}}\underline{\mathbf{h}}$, and we can express the MVDR beamformer alternatively as the minimum power distortionless response (MPDR) beamformer [30]. We obtain the MPDR beamformer interestingly by exploiting the correlation matrix of the noisy signals as

$$\underline{\mathbf{h}}_{\mathrm{MPDR-HB}} = \mathbf{R}_{\underline{\mathbf{y}}}^{-1}\underline{\mathbf{D}}_N(\theta_0,\omega_0)\left[\underline{\mathbf{D}}_N^H(\theta_0,\omega_0)\mathbf{R}_{\underline{\mathbf{y}}}^{-1}\underline{\mathbf{D}}_N(\theta_0,\omega_0)\right]^{-1}\mathbf{1}_N. \tag{A.48}$$

We can identify the harmonic model-based Wiener beamformer in (A.45) as the weighted MVDR beamformer in (A.47). The diagonal weight matrix $\mathbf{P}\left(\theta_0,\omega_0\right)$ is related to the narrowband input SNRs of the harmonics. Therefore, we can conclude that the MVDR and Wiener beamformers are approximately equivalent in high input SNRs. Moreover, it has also been shown

in [21] that we always have a trade-off in noise reduction and speech distortion index between the MVDR and Wiener beamformers, i.e.,

$$\begin{aligned} \text{oSNR}\left(\underline{\mathbf{h}}_{\text{W}-\text{HB}}\right) \geq \text{oSNR}\left(\underline{\mathbf{h}}_{\text{MVDR}-\text{HB}}\right) &\geq \text{iSNR}, && \text{(A.49)} \\ v_{\text{sd}}\left(\underline{\mathbf{h}}_{\text{W}-\text{HB}}\right) \geq v_{\text{sd}}\left(\underline{\mathbf{h}}_{\text{MVDR}-\text{HB}}\right) &= 0, && \text{(A.50)} \\ \xi_{\text{sr}}\left(\underline{\mathbf{h}}_{\text{W}-\text{HB}}\right) \geq \xi_{\text{sr}}\left(\underline{\mathbf{h}}_{\text{MVDR}-\text{HB}}\right) &= 1. && \text{(A.51)} \end{aligned}$$

6.3 Linearly Constrained Minimum Variance

We can derive a linearly constrained minimum variance (LCMV) beamformer [3, 31], which can handle more than one linear constraint, by exploiting the nullspace of the desired signal correlation matrix. Again, we assume the presence of a unique interference as explained in Subsection 5.2. The criterion to be optimized is now

$$\min_{\underline{\mathbf{h}}} \underline{\mathbf{h}}^H \mathbf{R}_{\underline{\mathbf{v}}} \underline{\mathbf{h}} \quad \text{subject to} \quad \underline{\mathbf{h}}^H \underline{\mathbf{C}} = \begin{bmatrix} \mathbf{1}_N^T & \mathbf{0}_{N_1}^T \end{bmatrix}, \tag{A.52}$$

where $\underline{\mathbf{C}}$ is defined in Subsection 5.2. We obtain

$$\underline{\mathbf{h}}_{\text{LCMV}-\text{HB}} = \mathbf{R}_{\underline{\mathbf{v}}}^{-1} \underline{\mathbf{C}} \left(\underline{\mathbf{C}}^H \mathbf{R}_{\underline{\mathbf{v}}}^{-1} \underline{\mathbf{C}} \right)^{-1} \begin{bmatrix} \mathbf{1}_N \\ \mathbf{0}_{N_1} \end{bmatrix}. \tag{A.53}$$

While the LCMV beamformer completely cancels the interference, there is no guarantee that the output SNR is greater than the input SNR [32]. The generalization of this LCMV beamformer to any number of interferences is straightforward, as long as the filter length ML is larger than the number of constraints. Now, we can express the linearly constrained minimum power (LCMP) beamformer, which utilizes the correlation matrix of the noisy signals, by the following equation:

$$\underline{\mathbf{h}}_{\text{LCMP}-\text{HB}} = \mathbf{R}_{\underline{\mathbf{y}}}^{-1} \underline{\mathbf{C}} \left(\underline{\mathbf{C}}^H \mathbf{R}_{\underline{\mathbf{y}}}^{-1} \underline{\mathbf{C}} \right)^{-1} \begin{bmatrix} \mathbf{1}_N \\ \mathbf{0}_{N_1} \end{bmatrix}. \tag{A.54}$$

Although the MVDR/LCMV and the MPDR/LCMP beamformers are theoretically the same, an inaccurate estimate of the correlation matrix in practice causes mismatch between the actual and the presumed signal in the MPDR/LCMP beamformers. Furthermore, the MVDR/LCMV beamformers are more robust to DOA estimation errors than the MPDR/LCMP beamformers [30, 33]. Therefore, for the sake of the maximum WNG, we can add the minimum filter norm constraint as $\underline{\mathbf{h}}^H \underline{\mathbf{h}} \leq k$ to the beamformers in addition to the distortionless constraints, where k is a positive constant. This modification corresponds to the so-called diagonal loading approach [11, 30, 34] which is given by $\mathbf{R}_{\underline{\mathbf{y}}} \leftarrow \mathbf{R}_{\underline{\mathbf{y}}} + \lambda \mathbf{I}_{ML}$, where λ is a positive constant. In general, the diagonal loading technique is applied to improve the performance of the beamformers with errors on the signal parameters (i.e., the DOA and frequency) and an inaccurate estimation of the correlation matrix.

6.4 Maximum SNR and Trade-Off

We designed the model-based beamformers based on the frequencies of periodic signals. Here, we unify the multichannel filtering approach in the subspace [24] with the harmonic signal model (A.15). The two Hermitian matrices $\mathbf{R}_{\underline{\mathbf{x}}}$ and $\mathbf{R}_{\underline{\mathbf{v}}}$ can be jointly diagonalized as follows [22]:

$$\mathbf{B}^H \mathbf{R}_{\underline{\mathbf{x}}} \mathbf{B} = \mathbf{\Lambda}, \tag{A.55}$$

$$\mathbf{B}^H \mathbf{R}_{\underline{\mathbf{v}}} \mathbf{B} = \mathbf{I}_{ML}, \tag{A.56}$$

where $\mathbf{B}$ is a full-rank square matrix (of size $ML \times ML$) and $\mathbf{\Lambda}$ is a diagonal matrix whose main elements are real and nonnegative. Furthermore, $\mathbf{\Lambda}$ and $\mathbf{B}$ are the eigenvalue and eigenvector matrices, respectively, of $\mathbf{R}_{\underline{\mathbf{v}}}^{-1} \mathbf{R}_{\underline{\mathbf{x}}}$, i.e.,

$$\mathbf{R}_{\underline{\mathbf{v}}}^{-1} \mathbf{R}_{\underline{\mathbf{x}}} \mathbf{B} = \mathbf{B} \mathbf{\Lambda}. \tag{A.57}$$

The eigenvalues of $\mathbf{R}_{\underline{\mathbf{v}}}^{-1} \mathbf{R}_{\underline{\mathbf{x}}}$ can be ordered as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N > \lambda_{N+1} = \cdots = \lambda_{ML} = 0$, and we denote the corresponding eigenvectors by $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_{ML}$. The correlation matrix (for consistency) can also be diagonalized as

$$\mathbf{B}^H \mathbf{R}_{\underline{\mathbf{y}}} \mathbf{B} = \mathbf{\Lambda} + \mathbf{I}_{ML}. \tag{A.58}$$

The maximum SNR beamformer is obtained by maximizing the output SNR. It is clear that (A.25) is maximized with

$$\underline{\mathbf{h}}_{\max} = \beta_1 \mathbf{b}_1, \tag{A.59}$$

where $\beta_1 \neq 0$ is an arbitrary complex number. The optimal value of β_1 is obtained by minimizing distortion. Substituting (A.59) into $J_{\mathrm{ds}}(\underline{\mathbf{h}})$ in (A.35) and minimizing the resulting expression with respect to β_1, we find the maximum SNR filter with minimum distortion:

$$\underline{\mathbf{h}}_{\max} = \frac{\mathbf{b}_1 \mathbf{b}_1^H}{\lambda_1} \mathbf{R}_{\underline{\mathbf{x}}} \mathbf{i}_{ML}. \tag{A.60}$$

Due to the relation $\mathbf{b}_1^H \mathbf{R}_{\underline{\mathbf{x}}} \mathbf{b}_1 = \lambda_1$ and $\mathbf{b}_1^H \mathbf{R}_{\underline{\mathbf{v}}} \mathbf{b}_1 = 1$, it can be verified that

$$\mathrm{oSNR}(\underline{\mathbf{h}}_{\max}) = \lambda_1, \tag{A.61}$$

which corresponds to the maximum output SNR, and

$$\mathrm{oSNR}(\underline{\mathbf{h}}) \leq \mathrm{oSNR}(\underline{\mathbf{h}}_{\max}), \ \forall \underline{\mathbf{h}}. \tag{A.62}$$

It is possible to derive a class of trade-off beamformers:

$$\underline{\mathbf{h}}_{\mathrm{T},Q} = \sum_{q=1}^{Q} \beta_q \mathbf{b}_q \tag{A.63}$$

$$= \sum_{q=1}^{Q} \frac{\mathbf{b}_q \mathbf{b}_q^H}{\lambda_q} \mathbf{R}_{\underline{\mathbf{x}}} \mathbf{i}_{ML}, \tag{A.64}$$

where $1 \leq Q \leq N$. We observe that for $Q = 1$ and $Q = N$, we obtain $\underline{\mathbf{h}}_{\mathrm{T},1} = \underline{\mathbf{h}}_{\max}$ and $\underline{\mathbf{h}}_{\mathrm{T},N} = \underline{\mathbf{h}}_{\mathrm{MVDR-HB}}$, respectively. We deduce that the output SNR and the speech reduction factor of the trade-off beamformer are, respectively, as [26]

$$\mathrm{oSNR}\left(\underline{\mathbf{h}}_{\mathrm{T},Q}\right) = \frac{\sum_{q=1}^{Q} \lambda_q |\beta_q|^2}{\sum_{q=1}^{Q} |\beta_q|^2}, \tag{A.65}$$

$$\xi_{\mathrm{sr}}\left(\underline{\mathbf{h}}_{\mathrm{T},Q}\right) = \frac{\mathbf{1}_N^T \mathbf{R}_{\mathbf{a}} \mathbf{1}_N}{\sum_{q=1}^{Q} \beta_q^2 \lambda_q}. \tag{A.66}$$

Therefore, we will have

$$\mathrm{oSNR}\left(\underline{\mathbf{h}}_{\mathrm{T},1}\right) \geq \mathrm{oSNR}\left(\underline{\mathbf{h}}_{\mathrm{T},2}\right) \geq \cdots \geq \mathrm{oSNR}\left(\underline{\mathbf{h}}_{\mathrm{T},N}\right) \tag{A.67}$$

and

$$\xi_{\mathrm{sr}}\left(\underline{\mathbf{h}}_{\mathrm{T},1}\right) \geq \xi_{\mathrm{sr}}\left(\underline{\mathbf{h}}_{\mathrm{T},2}\right) \geq \cdots \geq \xi_{\mathrm{sr}}\left(\underline{\mathbf{h}}_{\mathrm{T},N}\right). \tag{A.68}$$

7 Nonparametric Broadband Beamforming

Nonparametric broadband beamforming is a general technique without imposing a priori assumption regarding the signal. The general filter-and-sum beamformers, fixed and data-dependent beamformers, are designed over a wide frequency bands (see Fig. A.1). This is equivalent to decomposing the signal into K uniformly spaced frequency bands, i.e., $v_1, v_2, \ldots, v_K$, as

$$x(t) = \sum_{k=1}^{K} b_k e^{\jmath v_k [t - f_{\mathrm{s}} \tau_m(\theta)]}, \tag{A.69}$$

where $b_1, b_2, \ldots, b_K$ are complex spectral amplitudes of the corresponding frequencies. Therefore, the signal is modeled as

$$\underline{\mathbf{y}}(t) = \underline{\mathbf{D}}_K(\theta_0)\mathbf{b}(t) + \underline{\mathbf{v}}(t), \tag{A.70}$$

where $\mathbf{b}(t) = \left[\begin{array}{cccc} b_1 e^{\jmath v_1 t} & b_2 e^{\jmath v_2 t} & \cdots & b_K e^{\jmath v_K t} \end{array}\right]^T$ and

$$\underline{\mathbf{D}}_K(\theta_0) = \left[\begin{array}{ccc} \mathbf{d}_{1,1}(\theta_0, v_1) & \cdots & \mathbf{d}_{1,1}(\theta_0, v_K) \\ \mathbf{d}_{2,1}(\theta_0, v_1) & \cdots & \mathbf{d}_{2,1}(\theta_0, v_K) \\ \vdots & & \vdots \\ \mathbf{d}_{M,1}(\theta_0, v_1) & \cdots & \mathbf{d}_{M,1}(\theta_0, v_K) \end{array}\right]. \tag{A.71}$$

Such a broadband signal model leads to nonparametric broadband beamformers (BBs). The BB approach is a special case of the HB approach which

is based on the exact model of the signal in (A.2). The known BBs can be derived from either the nonparametric expression in (A.70), or the proposed HBs. For instance, the nonparametric broadband MVDR and Wiener beamformers, respectively, are given by

$$\underline{\mathbf{h}}_{\mathrm{MVDR-BB}} = \mathbf{R}_{\underline{\mathbf{v}}}^{-1}\underline{\mathbf{D}}_K(\theta_0)\Big[\underline{\mathbf{D}}_K^H(\theta_0)\mathbf{R}_{\underline{\mathbf{v}}}^{-1}\underline{\mathbf{D}}_K(\theta_0)\Big]^{-1}\mathbf{1}_K, \tag{A.72}$$

$$\underline{\mathbf{h}}_{\mathrm{W-BB}} = \mathbf{R}_{\underline{\mathbf{v}}}^{-1}\underline{\mathbf{D}}_K(\theta_0)\Big[\mathbf{R}_{\mathbf{b}}^{-1}+\underline{\mathbf{D}}_K^H(\theta_0)\mathbf{R}_{\underline{\mathbf{v}}}^{-1}\underline{\mathbf{D}}_K(\theta_0)\Big]^{-1}\mathbf{1}_K, \tag{A.73}$$

where $\mathbf{1}_K$ is the all ones column vector of length K. For $\underline{\mathbf{D}}_K^H(\theta_0)\mathbf{R}_{\underline{\mathbf{v}}}^{-1}\underline{\mathbf{D}}_K(\theta_0)$ to be invertible, we require that $K \leq L$. Substituting (A.72) and (A.73) into (A.25) and (A.31), for $N \leq K$, we get

$$\mathrm{oSNR}\left(\underline{\mathbf{h}}_{\mathrm{MVDR-HB}}\right) \geq \mathrm{oSNR}\left(\underline{\mathbf{h}}_{\mathrm{MVDR-BB}}\right), \tag{A.74}$$

$$v_{\mathrm{sd}}\left(\underline{\mathbf{h}}_{\mathrm{MVDR-BB}}\right) \geq v_{\mathrm{sd}}\left(\underline{\mathbf{h}}_{\mathrm{MVDR-HB}}\right) = 0, \tag{A.75}$$

and

$$\mathrm{oSNR}\left(\underline{\mathbf{h}}_{\mathrm{W-HB}}\right) \geq \mathrm{oSNR}\left(\underline{\mathbf{h}}_{\mathrm{W-BB}}\right), \tag{A.76}$$

$$v_{\mathrm{sd}}\left(\underline{\mathbf{h}}_{\mathrm{W-BB}}\right) \geq v_{\mathrm{sd}}\left(\underline{\mathbf{h}}_{\mathrm{W-HB}}\right) \geq 0. \tag{A.77}$$

These expressions show, interestingly, how the HBs are better than the BBs in noise reduction (these properties are valid for both the data-dependent and fixed beamformers).

8 Simulations

In this section, we give numerical examples to illustrate the performance of the proposed harmonic model-based beamformers (HBs) in different situations. We verify the relationship between the objective measures of the HBs, and compare them with the nonparametric broadband beamformers (BBs). The measures are the amount of noise reduction and the speech distortion index. The perceptual evaluation of speech quality (PESQ) [35] and short-time objective intelligibility measure (STOI) [36] are also used to evaluate the quality and intelligibility of the output speech signals.

8.1 Synthetic Signals

Firstly, we carried out simulations on synthetic signals generated using the harmonic signal model in (A.2) as well as computer-generated random noise (spatially white noise), and conducted Monte-Carlo simulations for different settings. The desired signal had the fundamental frequency $\omega_0 \in \mathcal{U}\{200, 500\} \times (2\pi/f_{\mathrm{s}})$, and $N = 5$ harmonics with complex amplitudes $|a_n| \in \mathcal{U}\{0.25, 1.5\}$

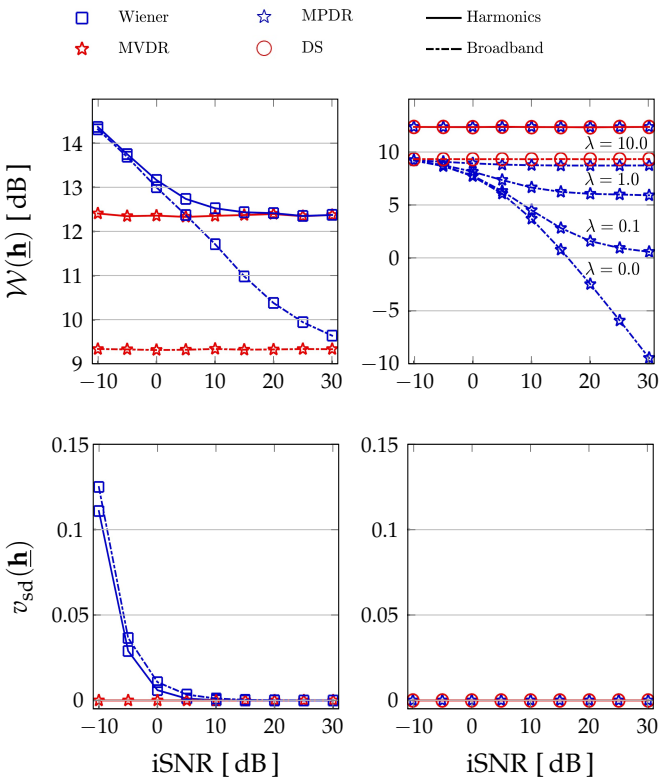


Fig. A.2: (Top) White noise gain and (bottom) speech distortion index of (left) the Wiener and MVDR beamformers and (right) the DS and MPDR beamformers as a function of the input SNR.

and phases $\phi_n \in \mathcal{U}\{-\pi, \pi\}$, where $\mathcal{U}$ denotes a random uniform distribution, and $f_{\mathrm{s}} = 8.0$ kHz is the sampling frequency. The signal has the DOA $\theta_0 \in \mathcal{U}\{0, \pi\}$ radians with respect to a uniform linear array (ULA) with the distance between two successive sensors $\delta = 0.04$ m, and the speed of sound $c = 343.2$ m/s. We assumed that the DOA, the fundamental frequency, and the number of corresponding harmonics are known in these experiments, though the parameters can be estimated using the methods of [14–17, 19, 20].

We investigate the relationship between the white noise gain (WNG), $\mathcal{W}(\underline{\mathbf{h}})$, and the speech distortion index, $v_{\mathrm{sd}}(\underline{\mathbf{h}})$, of the nonparametric and the harmonic model-based Wiener and MVDR beamformers versus different input signal-to-noise ratios (iSNRs). We applied $L = 30$ temporal and $M = 3$ spatial samples to design beamformers. The objective measures of the BBs and the HBs are plotted as dashed and solid lines, respectively, in the rest of the paper. The results in Fig. A.2(left) show that the Wiener beamformer

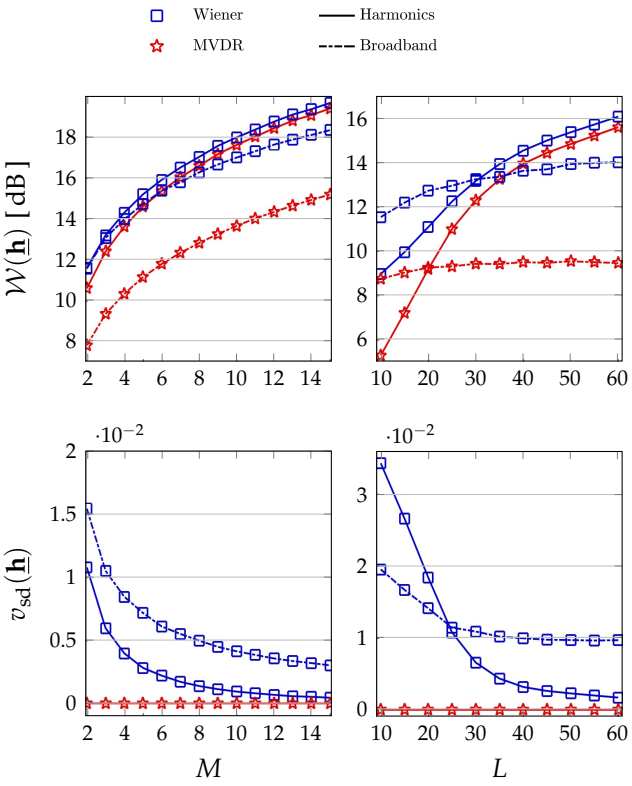


Fig. A.3: White noise gain and speech distortion index of the Wiener and MVDR beamformers as a function of the filter length.

reduces the noise while distorting the desired signal specifically in low input SNRs. The Wiener beamformer has a higher WNG than the MVDR beamformer, though it causes distortions on the speech signal. The WNG and the speech distortion index of the Wiener beamformer fall monotonically toward the MVDR beamformer when increasing the input SNR. In general, the MVDR-HB and W-HB obtain a higher WNG and lower speech distortion than the MVDR-BB and W-BB. Moreover, we compare the results of the MPDR-BB and MPDR-HB with the DS-BB and DS-HB, where the DS beamformer correspond to the MVDR beamformer in spatially white noise. Fig. A.2(right) shows the results of the MPDR beamformers using diagonal loading versus different λ and iSNR. We see that the MPDR-BB has the worst performance in high SNRs and with low diagonal loads. At the same time, the WNG and speech distortion index of the MPDR-HB, as well as the DS-BB and the DS-HB, are constant in different iSNRs.

In the next experiment, we evaluate the performance of the beamformers

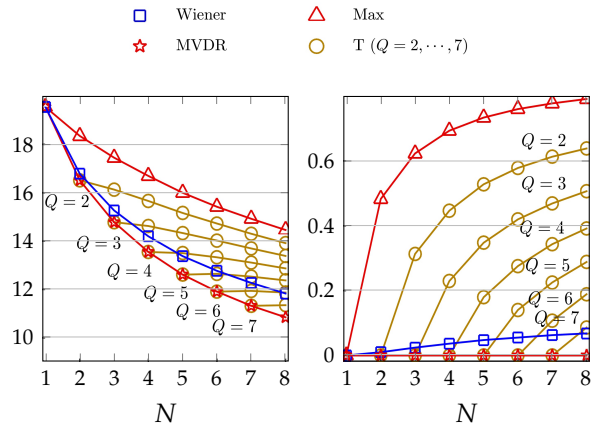


Fig. A.4: (Left) White noise gain, $\mathcal{W}(\underline{\mathbf{h}})$, in dB, and (right) speech distortion index, $v_{\mathrm{sd}}(\underline{\mathbf{h}})$, of the trade-off filter and the Wiener and MVDR harmonic model-based beamformers as a function of the number of harmonics.

versus the space-tapered filter length, ML, while keeping either $M = 3$ or $L = 30$. In this experiment, spatially white noise was added at 0 dB SNR. The results are depicted in Fig. A.3. As can be seen, the WNG increases with the filter length for the stationary signal. Therefore, a longer filter length should be applied for more noise reduction, and less distortion in the Wiener beamformers. Despite the WNG of the HBs directly relates to L, the WNG of BBs is fixed in large Ls. Therefore, the performances of HBs and BBs cross when varying L. In Fig. A.4, we can see that the performance of the Wiener-HB and the MVDR-HB is decreased with the number of harmonics, i.e., the HBs lose their degrees of freedom for a large number of constraints. We explore the properties of the trade-off filter for $Q = 1, \ldots, 7$. We recall that $Q = 1$ corresponds to the maximum SNR filter, and $Q = N$ corresponds to the MVDR-HB for the defined synthetic signal with N harmonics. More specifically, no distortion is achieved by exploiting as many significant eigenvectors as the number of harmonics. We can see that the performance of the trade-off filter depends on Q; the WNG and the speech distortion index of the trade-off filter increases when Q is decreased.

In the experiments, we considered no mismatch of the parameters of the model. Figure A.5 shows the normalized white noise gains of the DS-BB and DS-HB in the presence of DOA and fundamental frequency mismatches. The beamformers $\underline{\mathbf{h}}$ and $\underline{\tilde{\mathbf{h}}}$ were designed, respectively, using the true parameters, (θ_0, ω_0), and the mismatched parameters, $(\hat{\theta}_0, \hat{\omega}_0)$. In this experiment, we applied $M = 3$ and $L = 80$, and the signal source had $\theta_0 = \pi/2$, $\omega_0 = 200 \times (2\pi/f_{\mathrm{s}})$, and $N = 1, 2, \ldots, 5$. As can be seen, the gain of the DS-HB is sensitive to mismatch of the DOA and the fundamental frequency, and it is

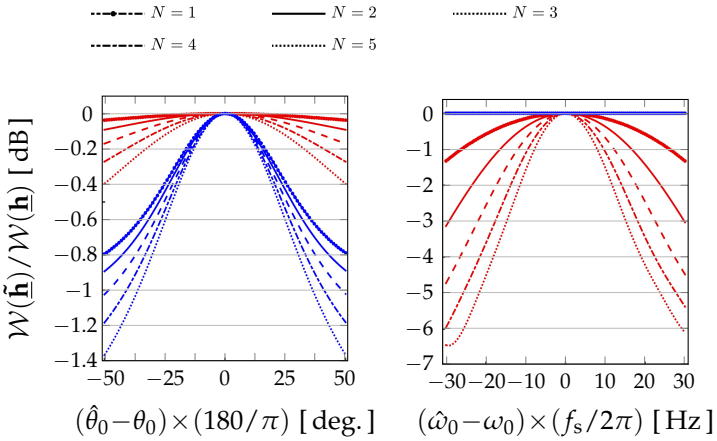
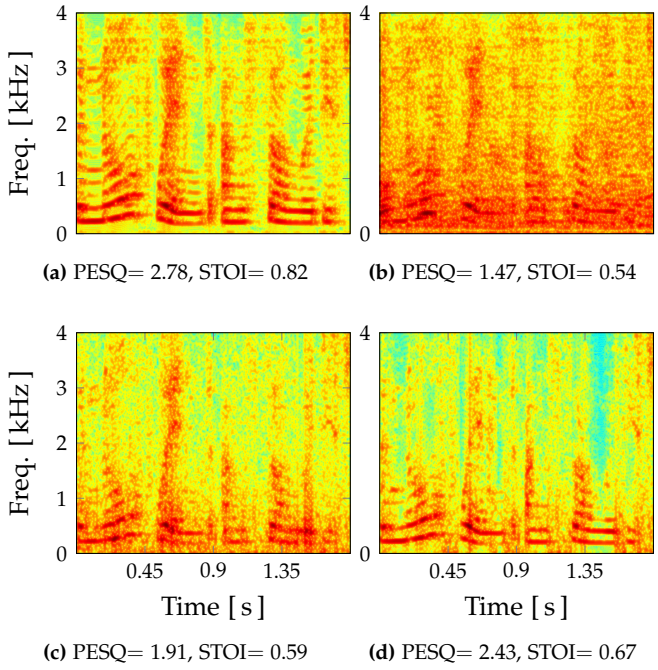


Fig. A.5: Normalized white noise gain of the harmonic model-based (red color) and the broadband (blue color) beamformerers in the presence of the DOA and the fundamental frequency mismatch, respectively.

less sensitive to DOA mismatch than the DS-BB.

8.2 Real-Life Experiments

In this section, we evaluate the proposed beamformers in real-life scenarios. We considered a reverberant environment, and for generating the multichannel signals, we employed an online MATLAB implementation [37] of the image method [38] with the maximum amount of sound reflections from the reflecting walls in the room impulse response (RIR) generator [37]. We simulated the room size $6 \times 5 \times 3$ m (length$\times$width$\times$height), and placed two simultaneous speakers (one male and one female) at the fixed locations $2 \times 4 \times 1.5$ m and $4 \times 4 \times 1.5$ m that results in directions of arrival $\theta_0 = 71.6^\circ$ and $\theta_1 = 108.4^\circ$ with respect to the array centered at $3 \times 1 \times 1.5$ m. The uniform array included $M = 4$ microphones with the same distance δ as in the previous experiment. The reflection coefficient was set to achieve $T_{60} = 200$ ms reverberation time. The microphone outputs were simulated by convolving the speech signals with the RIRs, and added to multichannel and spatially coherent babble noise [39]. We employed the Keele database [40], which consists of male and female speech, and downsampled the speech signals with the sampling frequency $f_s = 8.0$ kHz. The interfering speech signal level was as hight as the desired signal, i.e., 0 dB signal-to-interference ratio (SIR), and the babble noise was added at 5 dB SNR. We designed the beamformers using the estimated parameters of the voiced speech parts. We estimated the fundamental frequency and the number of harmonics from the separate clean nonreverberant signals (direct-path signals) using frames of samples.



(a) PESQ= 2.78, STOI= 0.82 **(b)** PESQ= 1.47, STOI= 0.54

(c) PESQ= 1.91, STOI= 0.59 **(d)** PESQ= 2.43, STOI= 0.67

Fig. A.6: Spectrograms of (a) the reverberant original signal, (b) the noisy signal (mixed at 5 dB SNR of the babble noise and 0 dB SIR of an interfering speech signal), and (c-d) the output signals of the LCMV-BB and the LCMV-HB.

The frames' size were 20 ms (corresponding to 160 temporal samples) and updated every 10 ms. For speech processing, 20–30 ms of voiced speech signals is commonly assumed as a short-term stationary signal. We applied the non-linear least-squares (NLS) fundamental frequency estimator [15] and the Bayesian information criterion (BIC) model order estimator [41]. Moreover, we applied the normalized low frequency energy ratio (NLFER) in the frequency domain [42] to select the frames which contain voiced speech. We selected the frames having an $\text{NLFER}(t) \geq 0.4$ over the frequency range $[60\ ,\ 420] \times (2\pi/f_s)$ radians per second, where $\text{NLFER}(t)$ is the ratio of the signal energy over the entire signal. The beamformers had $L = 60$ taps and updated every 10 ms. In order to estimate the noise statistics, a voice activity detection (VAD) algorithm [43, 44] can be used to recognize non-speech frames in the utterance. However, in the presence of non-stationary noise and competing speakers, VAD is not trivial. Hence, in this experiment, we estimated the correlation matrix $\hat{\mathbf{R}}_{\mathbf{v}}$ directly from the noise signal based on

recursive averaging:

$$\hat{\mathbf{R}}_{\mathbf{v}}(t) = \alpha \hat{\mathbf{R}}_{\mathbf{v}}(t-T) + \frac{1-\alpha}{T} \sum_{s=0}^{T-1} \underline{\mathbf{v}}(t-s)\underline{\mathbf{v}}^H(t-s), \tag{A.78}$$

and regularized the correlation matrix [45] such that

$$\hat{\mathbf{R}}_{\mathbf{v},\text{reg}}(t) = (1-\gamma)\hat{\mathbf{R}}_{\mathbf{v}}(t) + \gamma \frac{\text{tr}\left[\hat{\mathbf{R}}_{\mathbf{v}}(t)\right]}{ML} \mathbf{I}_{ML}. \tag{A.79}$$

We chose $T = 100$, $\alpha = 0.2$, and $\gamma = 0.001$ to the best results in terms of output SNR and perceptual score. We retrieved the correlation matrix of the clean voiced speech directly from the corresponding parameter estimates as $\hat{\mathbf{R}}_{\mathbf{x}} = \underline{\mathbf{D}}_{\hat{N}}(\theta_0, \hat{\omega}_0)\hat{\mathbf{R}}_{\mathbf{a}}\underline{\mathbf{D}}_{\hat{N}}^H(\theta_0, \hat{\omega}_0)$ with $\hat{\mathbf{R}}_{\mathbf{a}} = \text{diag}\begin{pmatrix} |\hat{a}_1|^2 & |\hat{a}_2|^2 & \cdots & |\hat{a}_N|^2 \end{pmatrix}$. The complex amplitudes were estimated using the least-squares estimator. For the frames of unvoiced signals, we applied the corresponding BBs.

In the following experiment, the spectrogram of 1.8 seconds of the reverberant speech signal and its noisy mixture with the babble noise and interference are shown in Figs. A.6(a-b). The output signals of the nonparametric and harmonic model-based LCMV beamformers are shown, respectively, in Figs. A.6(c-d). As can be seen, the LCMV-HB have less noise between the harmonics than the LCMV-BB. In this experiment, the PESQ and STOI scores were measured from differences between the direct-path clean speech signal and the filtered signals. At the bottom of the spectrograms, the measures show that the examined HB is closer to the original speech signal than the corresponding BB. We evaluate the performance of the tailored beamformers as a function of the noise level. The simulation results in Figs. A.7 shows that all HBs perform well, achieving a higher output SNR in addition to higher PESQ and STOI scores than the corresponding BBs. For example, although the null forming nonparametric beamformer (NF-BB) has lower performance measures than the noisy input signal, the corresponding harmonic model-based beamformer NF-HB has better results. The maximum SNR filter achieves the maximum output SNR, but its PESQ and STOI scores are lower than the other model-based beamformers. We see from the figures that the output SNR of the beamformers increases with the input SNR, similar to the previous experiments on the synthetic signal. Moreover, although the Wiener-HB has a higher output SNR and PESQ score than the MVDR-HB, we can see that the MVDR-HB is more intelligible than the Wiener-HB.

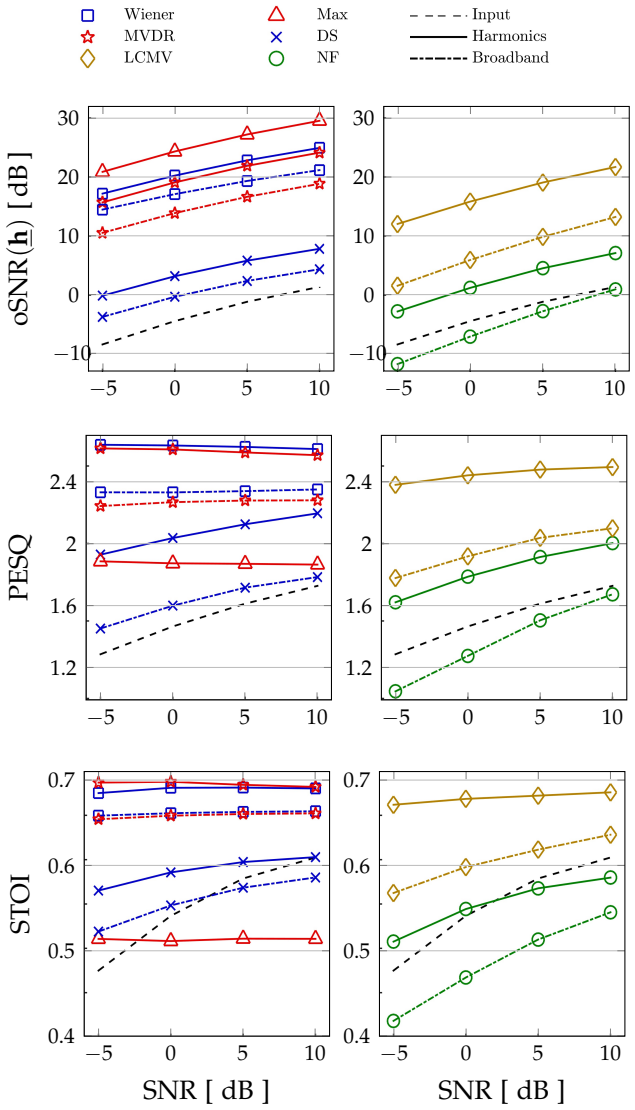


Fig. A.7: According to the order of plots, output SNR, PESQ, and STOI of the beamformers as a function of the SNR of the babble noise in a real-life experiment.

In the last experiment, we conducted the previous experiment setup in 5 dB SNR and without the interfering speech source. Figure A.8 compares the PESQ and the STOI measures of the enhanced and noisy signals in different reverberation times. As expected, the DS-HB and MVDR-HB perform better than the corresponding BBs. As can be seen, the DS-HB is also more intelligible than the MVDR-BB.

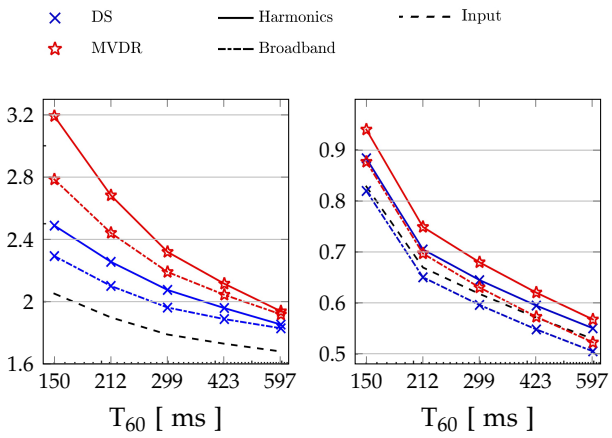


Fig. A.8: (Left) PESQ and (right) STOI scores of the nonparametric and harmonic model-based beamformers as a function of the reverberation time.

9 Conclusion

In this paper, we have presented a new class of broadband beamformers. In the proposed technique, we have exploited a priori knowledge about voiced speech signals to develop model-based beamforming. We started by formulating multichannel signals considering the spatial and the spectral properties of periodic signals. We have exploited the harmonic frequencies that give us an advantage to decompose harmonic sources with regard to respective fundamental frequencies. Experiments on synthetic and real signals have demonstrated the properties of the proposed harmonic model-based beamformers, compared with nonparametric beamformers. The most important observation from the experiments is that the harmonic model-based beamforming is superior to the nonparametric beamforming in speech enhancement with higher quality and intelligibility scores.

References

[1] J. Benesty, Y. Huang, and J. Chen, *Microphone Array Signal Processing*. Springer-Verlag, 2008, vol. 1.

[2] J. Flanagan, J. Johnston, R. Zahn, and G. Elko, "Computer-steered microphone arrays for sound transduction in large rooms," *J. Acoust. Soc. Am.*, vol. 78, no. 5, pp. 1508–1518, 1985.

[3] O. L. Frost, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, vol. 60, no. 8, pp. 926–935, Aug. 1972.

[4] B. D. Van Veen and K. M. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE ASSP Mag.*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[5] M. Brandstein and D. Ward, Eds., *Microphone Arrays - Signal Processing Techniques and Applications*. Springer-Verlag, 2001.

[6] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Signal Process.*, vol. 49, no. 8, pp. 1614–1626, 2001.

[7] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.

[8] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays*. Springer, 2001, pp. 39–60.

[9] A. Nehorai and B. Porat, "Adaptive comb filtering for harmonic signal enhancement," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 5, pp. 1124–1138, Oct. 1986.

[10] T. Nakatani, M. Miyoshi, and K. Kinoshita, "Single-microphone blind dereverberation," in *Speech Enhancement*. Springer, 2005, pp. 247–270.

[11] R. T. Lacoss, "Data Adaptive Spectral Analysis Methods," *Geophysics*, vol. 36, pp. 661–675, Aug. 1971.

[12] M. G. Christensen and A. Jakobsson, "Optimal filter designs for separating and enhancing periodic signals," *IEEE Trans. Signal Process.*, vol. 58, no. 12, pp. 5969–5983, Dec. 2010.

[13] J. R. Jensen, M. G. Christensen, J. Benesty, and S. H. Jensen, "Joint spatio-temporal filtering methods for DOA and fundamental frequency estimation," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 23, no. 1, pp. 174–185, 2015.

[14] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 111, no. 4, pp. 1917–1930, Apr. 2002.

[15] M. G. Christensen and A. Jakobsson, "Multi-pitch estimation," *Synthesis Lectures on Speech and Audio Process.*, vol. 5, no. 1, pp. 1–160, 2009.

[16] P. Stoica and Y. Selen, "Model-order selection: a review of information criterion rules," *IEEE Signal Process. Mag.*, vol. 21, no. 4, pp. 36–47, Jul. 2004.

[17] M. S. Brandstein and H. Silverman, "A practical methodology for speech source localization with microphone arrays," *Comput. Speech Language*, 1997.

[18] S. Karimian-Azari, J. R. Jensen, and M. G. Christensen, "Computationally efficient and noise robust DOA and pitch estimation," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 24, no. 9, pp. 1609–1621, 2016.

[19] ——, "Fast joint DOA and pitch estimation using a broadband MVDR beamformer," in *Proc. European Signal Processing Conf.*, Sept. 2013, pp. 1–5.

[20] J. R. Jensen, M. G. Christensen, and S. H. Jensen, "Nonlinear least squares methods for joint DOA and pitch estimation," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 21, no. 5, pp. 923 –933, May 2013.

[21] J. Benesty, J. Chen, and Y. Huang, "Speech enhancement in the karhunen-loève expansion domain," *Synthesis Lectures on Speech and Audio Processing*, vol. 7, no. 1, pp. 1–112, 2011.

[22] J. N. Franklin, *Matrix Theory*. Prentice-Hall, 1968.

[23] Y. Lacouture-Parodi, E. A. Habets, J. Chen, and J. Benesty, "Multichannel noise reduction in the karhunen-loève expansion domain," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 22, no. 5, pp. 923–936, 2014.

[24] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Trans. Signal Process.*, vol. 50, no. 9, pp. 2230–2244, Sep. 2002.

[25] Y. Hu and P. C. Loizou, "A subspace approach for enhancing speech corrupted by colored noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1. IEEE, 2002, pp. I–573.

[26] J. R. Jensen, J. Benesty, and M. G. Christensen, "Noise reduction with optimal variable span linear filters," *IEEE Trans. Acoust., Speech, Signal Process.*, 2016.

[27] M. G. Christensen, "Accurate estimation of low fundamental frequencies from real-valued measurements," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 21, no. 10, pp. 2042–2056, 2013.

[28] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction Wiener filter," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, no. 4, pp. 1218–1234, Jul. 2006.

[29] J. Benesty, J. Chen, Y. A. Huang, and S. Doclo, "Study of the wiener filter for noise reduction," in *Speech Enhancement*. Springer, 2005, pp. 9–41.

[30] H. Cox, "Resolving power and sensitivity to mismatch of optimum array processors," *J. Acoust. Soc. Am.*, vol. 54, no. 3, pp. 771–785, Sep. 1973.

[31] M.-H. Er and A. Cantoni, "Derivative constraints for broad-band element space antenna array processors," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 31, no. 6, pp. 1378–1393, 1983.

[32] M. Souden, J. Benesty, and S. Affes, "A study of the LCMV and MVDR noise reduction filters," *IEEE Trans. Signal Process.*, vol. 58, no. 9, pp. 4925–4935, Sept. 2010.

[33] L. Ehrenberg, S. Gannot, A. Leshem, and E. Zehavi, "Sensitivity analysis of MVDR and MPDR beamformers." IEEE, 2010, pp. 416–420.

[34] B. D. Carlson, "Covariance matrix estimation errors and diagonal loading in adaptive arrays," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 24, no. 4, pp. 397–401, 1988.

[35] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, 2001, pp. 749–752 vol.2.

[36] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2010, pp. 4214–4217.

[37] E. A. P. Habets, "Room impulse response generator," Technische Universiteit Eindhoven, Eindhoven, Netherlands, Tech. Rep., 2010, ver. 2.0.20100920.

[38] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[39] E. A. P. Habets, I. Cohen, and S. Gannot, "Generating nonstationary multisensor signals under a spatial coherence constraint," *The Journal of the Acoustical Society of America*, vol. 124, no. 5, pp. 2911–2917, 2008.

[40] F. Plante, G. F. Meyer, and W. A. Ainsworth, "A pitch extraction reference database," in *Proc. Eurospeech*, Sep. 1995, pp. 837–840.

[41] P. Djuric, "A model selection rule for sinusoids in white gaussian noise," *IEEE Trans. Signal Process.*, vol. 44, no. 7, pp. 1744–1751, Jul 1996.

[42] S. A. Zahorian and H. Hu, "A spectral/temporal method for robust fundamental frequency tracking," *J. Acoust. Soc. Am.*, vol. 123, no. 6, pp. 4559–4571, 2008.

[43] J. Ramírez, J. C. Segura, C. Benítez, L. García, and A. Rubio, "Statistical voice activity detection using a multiple observation likelihood ratio test," *IEEE Signal Process. Lett.*, vol. 12, no. 10, pp. 689–692, 2005.

[44] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, 1999.

[45] F. van der Heijden, R. P. W. Duin, D. de Ridder, and D. M. J. Tax, *Classification, Parameter Estimation and State Estimation - An Engineering Approach using MATLAB®*. John Wiley & Sons Ltd, 2004.

Paper B

Computationally Efficient and Noise Robust DOA and Pitch Estimation

Sam Karimian-Azari, Jesper Rindom Jensen,
and Mads Græsbøll Christensen

The paper has been published in the
IEEE Transactions on Audio, Speech and Language Processing, 2016.

Abstract

Many natural signals, such as voiced speech and some musical instruments, are approximately periodic over short intervals. These signals are often described in mathematics by the sum of sinusoids (harmonics) with frequencies that are proportional to the fundamental frequency, or pitch. In sensor (microphone) array signal processing, the periodic signals are estimated from spatio-temporal samples regarding to the direction of arrival (DOA) of the signal of interest. In this paper, we consider the problem of pitch and DOA estimation of quasi-periodic audio signals. In real life scenarios, recorded signals are often contaminated by different types of noise, which challenges the assumption of white Gaussian noise in most state-of-the-art methods. We establish filtering methods based on noise statistics to apply to nonparametric spectral and spatial parameter estimates of the harmonics. We design minimum variance solutions with distortionless constraints to estimate the pitch from the frequency estimates, and to estimate the DOA from multichannel phase estimates of the harmonics. Applying this filtering method as the sum of weighted frequency and DOA estimates of the harmonics, we also design a joint DOA and pitch estimator. In white Gaussian noise, we derive even more computationally efficient solutions which are designed using the narrowband power spectrum of the harmonics. Numerical results reveal the performance of the estimators in colored noise compared with the Cramér-Rao lower bound. Experiments on real-life signals indicate the applicability of the methods in practical low local signal-to-noise ratios.

1 Introduction

Audio communication systems, such as teleconferencing, hearing-aids, and telecommunications, receive audio signals along with interferences and noise that degrade the quality and intelligibility (for speech signals) of the signals of interest. Audio source separation and enhancement are therefore relevant but challenging problems in audio and speech signal processing, and many works have already been devoted to them [1]. Filtering methods are common solutions for the enhancement problem, and recent parametric approaches rely on the periodic signal model [2–4]. Therefore, accurate parameter estimates of the periodic signals are required. The basic idea is that some audio signals, such as voiced speech and harmonic musical instruments, are approximately periodic over short intervals, and Fourier series describes such a signal as the sum of sinusoids (harmonics), which have frequencies proportional to the fundamental frequency, or *pitch* as it is commonly referred to. The harmonic signal model is exploited in many pitch estimation methods [5], e.g., the subspace orthogonality based method [5, 6] and the Markov-like weighted least squares (WLS) pitch estimator [7]. The WLS pitch estimator is the computationally efficient solution with good statistical performance

in white Gaussian noise. However, either missing or spurious frequency estimates in practice, often result in large pitch estimation errors [5].

Microphone array signal processing methods provide tools to improve audio communication systems. They utilize spatio-temporal samples to separate audio signals coming from different directions, relative to the array, at the same time. For example, a beamformer can steer the array outputs in the direction of the signal of interest (interested readers can find an overview of existing methods in [8] and [9]). The estimation of the direction of arrival (DOA) is a crucial and challenging problem, especially in noisy environments [10]. A general DOA estimator is first to estimate the time-difference of arrival (TDOA) between the microphones, e.g., using the correlation [11–13] and the phase shift [14] estimation methods, and then map the TDOA estimates to a DOA estimate. Although the conventional TDOA estimators are designed with a single source assumption [10], they possess an advantage over the other DOA estimation methods in terms of computational complexity [10]. For multiple signal sources, the broadband TDOA estimation method [14] has been used in the time-frequency domain [15], and the harmonic signal model has been used recently in [16]. The harmonic model-based DOA estimator [16] has been designed based on the WLS method by exploiting the weight matrices which are the Fisher Information Matrices (FIMs) in spatial white Gaussian noise. The estimator attains the Cramér-Rao lower bound (CRLB), and outperforms such state-of-the-art methods as the steered response power (SRP) method [17] and the position-pitch plane based (POPI) method [18].

In real-life scenarios, audio signals are often recorded in the presence of different types of background noise, e.g., traffic noise and wind noise, that have non-uniform power over the entire spectrum [19]. Therefore, the performance of the pitch and the broadband DOA estimation methods is degraded because of the incorrect white Gaussian noise assumption. However, an ideal estimator should be robust against different noise types. Furthermore, the vast majority of the pitch and the DOA estimators have traditionally been designed separately and applied in a sequential process, i.e., by estimating either the DOA of the given pitch or estimating the pitch of the signal received from the known DOA. Using a sequential process in the case of multiple signal sources with overlapping properties, either spatial or temporal parameters, the other parameter may not be estimated correctly. For example, if two sources at different angles have the same pitch, an estimation of their DOA is not trivial if the pitch is estimated first. Therefore, joint estimation of DOA and pitch would be beneficial for both estimates. Some methods have recently been proposed based on the two-dimensional (2D) spectral density estimation that intuitively characterizes spectral and spatial properties of the periodic signals [20]. We can divide the estimators generally into three groups of methods: the subspace based methods [21, 22], the filtering

based methods [23–25], and the statistical based methods, e.g., the nonlinear least squares (NLS) method [26]. The subspace and the filtering based methods apply the correlation matrix estimate of the spatio-temporal signals. Although the recently proposed joint spatio-temporal filtering method [25] is an optimal solution based on the Capon spectral density estimator, it has the disadvantage of a high computational complexity on large spatio-temporal signals. In white noise, the joint spatio-temporal filter expression is the same as the NLS joint estimator [26]. Moreover, for the most real world signals, the angle and the frequency of periodic signals are continuous parameters that correspond to highly coherent spectral density with large angle and frequency grids. In order to obtain high resolution estimates in many of the joint DOA and pitch estimation methods, a high computational complexity would be required, since they require two-dimensional grid searches that limits the feasibility of the methods in real-time applications.

In this paper, we propose consistent and computationally efficient solutions concerning the properties of the model of periodic signals in the presence of Gaussian noise. We estimate the fundamental frequency of periodic signals from unconstrained frequency estimates (UFEs) of the harmonics considering the fact that the frequency estimates of the harmonics may have different uncertainties over the spectrum [27, 28]. The UFEs have different uncertainties related to the reciprocal of narrowband signal-to-noise ratio (SNR) of the harmonics. Therefore, we can estimate the properties of noise at the frequencies of the harmonics from statistics of the UFEs. We propose a filtering solution that resembles the sum of weighted UFEs, and design a minimum variance distortionless response (MVDR) filter as an optimal solution to estimate the pitch from the UFEs. For the DOA estimation problem, we consider the properties of spatio-temporal signals, and design two filters based on the two-step procedure in [16]. First, we consider the linear relationship between the unwrapped multichannel phase estimates of the harmonics which are related to the so–called spatial frequencies, and design a filter to estimate the spatial frequencies of the harmonics. We then estimate the DOA from the estimated spatial frequencies of the harmonics. In each of these two steps, we estimate the properties of noise from statistics of the multichannel phase and DOA estimates of the harmonics. Moreover, we extend the proposed estimators into a joint pitch and DOA estimator from unconstrained frequency and DOA estimates of the harmonics using a 2D spectral density of spatio-temporal signals. Finally, we formulate simplified yet rigorous estimators in white noise. We derive the WLS pitch and DOA estimators in [7] and [16], respectively, from the proposed optimal approaches, and also propose a joint DOA and pitch estimator from the designed optimal filter in white noise as the computationally efficient solution.

The proposed estimators are computationally simple which resemble the sum of weighted parameters (i.e., the frequencies, the phases, and the DOAs)

of the harmonics. However, the most state-of-the-art parametric DOA and pitch estimators search over the signals that makes them computationally complex. Moreover, we estimate the weights using the properties of the signal and noise. Therefore, the proposed estimators are robust against different noise types.

The remainder of this paper is organized as follows: in the next section we present multichannel harmonic signal model. We present the pitch and the DOA estimation methods separately in Sections 3 and 4 and the joint estimation method in Section 5. Next, we formulate the CRLB of the parameter estimates in Section 6. We conduct some experiments, and analyze the computational complexity of the methods in Section **??**. We then conclude this work in Section 7.

2 Signal Model

2.1 Single-Channel Signal Model

Harmonic signals are modeled as the sum of sinusoids. We exploit discrete-time analytical signals to simplify the notation and facilitate a fast implementation, as obtained using the methods detailed in [29, 30]. We therefore define such a signal as L harmonics with frequencies $\omega_l \in [0, \pi]$, real amplitudes α_l, and phases $\psi_l \in [-\pi, \pi]$ for $l = 1, \ldots, L$ at the discrete-time index n as

$$x_0(n) = \sum_{l=1}^{L} \alpha_l \, e^{j\,(\omega_l n + \psi_l)}, \tag{B.1}$$

where $j = \sqrt{-1}$. The frequencies of the harmonics are integer multiples of the fundamental frequency ω_0 such that

$$\boldsymbol{\Omega} \triangleq \begin{bmatrix} \omega_1 & \omega_2 & \ldots & \omega_L \end{bmatrix}^T = \mathbf{d}_L \, \omega_0, \tag{B.2}$$

where $\mathbf{d}_L = \begin{bmatrix} 1 & 2 & \ldots & L \end{bmatrix}^T$, and the superscript ^{T} is the transpose operator. We assume that the observed signal $x_0(n)$ is contaminated with the complex-valued Gaussian noise $v_0(n)$ with zero mean, i.e.,

$$y_0(n) = x_0(n) + v_0(n). \tag{B.3}$$

The real and imaginary parts of the noise are uncorrelated and have the equal narrowband power spectrum $\Phi_0(\omega)/2$, for $\omega \in [0, 2\pi]$. At a high narrowband SNR at the frequency ω_l, i.e., $\text{SNR}^l = \alpha_l^2 / \Phi_0(\omega_l) \gg 1$, we consider that the phase of each sinusoid is perturbed by the real phase-noise $\Delta\psi_l(n)$, given by [27]

$$\Delta\psi_l(n) = \frac{|v_0(n)|}{\alpha_l} \sin(l\omega_0 n + \psi_l + \varrho_0), \tag{B.4}$$

where ϱ_0 is a uniformly distributed random phase on $[-\pi, \pi]$ [31]. The phase-noise causes local oscillations from the center of $l\omega_0 n + \psi_l$ with the variance given by

$$\mathrm{E}\{\Delta\psi_l^2(n)\} = \frac{1}{2\,\mathrm{SNR}^l} = \frac{\Phi_0(\omega_l)}{2\,\alpha_l^2}, \tag{B.5}$$

where $\mathrm{E}\{\cdot\}$ denotes the statistical expectation. It can be shown that the phase-noise of the harmonics are uncorrelated, i.e., $\mathrm{E}\{\Delta\psi_i(n)\Delta\psi_k(n)\} = 0$, for $i \neq k$, and the covariance matrix of the phase-noise vector $\Delta\boldsymbol{\Psi}_0 = \begin{bmatrix}\Delta\psi_1(n) & \Delta\psi_2(n) & \dots & \Delta\psi_L(n)\end{bmatrix}^T$ is

$$\begin{aligned}\mathbf{R}_{\Delta\boldsymbol{\Psi}_0} &= \mathrm{E}\{\Delta\boldsymbol{\Psi}_0\Delta\boldsymbol{\Psi}_0^T\} \\ &= \frac{1}{2}\,\mathrm{diag}\Big\{\frac{\Phi_0(\omega_1)}{\alpha_1^2} \;\; \frac{\Phi_0(\omega_2)}{\alpha_2^2} \;\; \dots \;\; \frac{\Phi_0(\omega_L)}{\alpha_L^2}\Big\},\end{aligned} \tag{B.6}$$

where $\mathrm{diag}\{\cdot\}$ denotes the diagonal matrix formed with the vector input along its diagonal. We therefore approximate the noisy signal by converting the additive noise to the phase-noise:

$$y_0(n) \approx \sum_{l=1}^{L} \alpha_l\, e^{j\,(\omega_l n + \psi_l + \Delta\psi_l(n))}. \tag{B.7}$$

This expression shows that the additive noise signal, in high narrowband SNRs, distorts only the phases of the sinusoids.

2.2 Multichannel Signal Model

We consider M omnidirectional microphones of an array receive a plane wave from a far field harmonic source at the direction $\theta_0 \in [-\pi/2, \pi/2]$ radians in an anechoic environment. We assume the first microphone as the reference, and model the signal at the m+1th microphone, for $m = 0, 1, \dots, M-1$, as a delayed signal due to the model of the array as $x_m(n) = x_0(n - \mathcal{F}_m)$, where $\mathcal{F}_m$ is the relative delay between the first and the $m+1$th microphone. Hence, the array output is given by

$$\begin{aligned}\mathbf{x}(n) &= \begin{bmatrix}x_0(n) & x_1(n) & \dots & x_{M-1}(n)\end{bmatrix}^T \\ &= \sum_{l=1}^{L} \alpha_l\, e^{j\,(\omega_l n + \psi_l)}\, \mathbf{z}_{\mathrm{s}}(\omega_{l,\mathrm{s}}),\end{aligned} \tag{B.8}$$

where $\mathbf{z}_{\mathrm{s}}(\omega_{l,\mathrm{s}})$ is called the steering vector that causes different discrete phase shifts among the microphones as a function of the spatial frequency $\omega_{l,s}$. Here we exploit a uniform linear array (ULA) to prove the concept, but the results

can be generalized for different structures of the array. Hence, the relative delay of the ULA is given by $\mathcal{F}_m = m f_s \tau_0 \sin(\theta_0)$, and the steering vector that corresponds to the hypothesized plane wave of the lth harmonic is

$$\mathbf{z}_s(\omega_{l,s}) = \begin{bmatrix} 1 & e^{-j\omega_{l,s}} & \dots & e^{-j(M-1)\omega_{l,s}} \end{bmatrix}^T, \tag{B.9}$$

where $\omega_{l,s} = \omega_l f_s \tau_0 \sin(\theta_0)$ with sampling frequency f_s and delay $\tau_0 = \delta/c$ between two adjacent sensors with a distance of δ and speed of sound c. We assume that the multichannel signals are contaminated with the additive complex-valued Gaussian noise $v_m(n)$, i.e.,

$$y_m(n) = x_m(n) + v_m(n), \text{ for } m = 0, 1, \dots, M-1, \tag{B.10}$$

with zero mean and the narrowband power spectrum $\Phi_m(\omega)/2$. At a high narrowband SNR, $\text{SNR}_m^l = \alpha_l^2/\Phi_m(\omega_l) \gg 1$, the additive Gaussian noise in each channel can be converted into an equivalent phase-noise, i.e.,

$$\Delta\psi_{l,m}(n) = \frac{|v_m(n)|}{\alpha_l} \sin(\omega_l n + \psi_l - \omega_{l,s} m + \varrho_m), \tag{B.11}$$

where ϱ_m is a uniformly distribute random phase $[-\pi, \pi]$, and the covariance matrix of the phase-noise vector $\Delta\mathbf{\Psi}_m = \begin{bmatrix} \Delta\psi_{1,m}(n) & \Delta\psi_{2,m}(n) & \dots & \Delta\psi_{L,m}(n) \end{bmatrix}^T$ is given by

$$\mathbf{R}_{\Delta\mathbf{\Psi}_m} = \frac{1}{2} \text{diag}\Big\{ \frac{\Phi_m(\omega_1)}{\alpha_1^2} \;\; \frac{\Phi_m(\omega_2)}{\alpha_2^2} \;\; \dots \;\; \frac{\Phi_m(\omega_L)}{\alpha_L^2} \Big\}. \tag{B.12}$$

Finally, we approximate the noisy signal as

$$y_m(n) \approx \sum_{l=1}^{L} \alpha_l \, e^{j(\omega_l n + \psi_l - \omega_{l,s} m + \Delta\psi_{l,m}(n))}. \tag{B.13}$$

3 Pitch Estimation

Spectral density estimate describes and analyzes a signal, and a maximum-likelihood (ML) estimator of a single sinusoid finds the peak location of the spectral density [32]. Over the N most recent time samples of a stationary signal, i.e., $\mathbf{y}_m(n) = \begin{bmatrix} y_m(n) & y_m(n+1) & \dots & y_m(n+N-1) \end{bmatrix}^T$, the variance of the ML frequency estimator is related to the reciprocal of the cubic number of the samples and the narrowband signal-to-noise ratio (SNR) [33]. We can consider enough samples to resolve closely spaced multiple sinusoids [34]. Moreover, the estimation methods of the other parameters of the given frequencies have been investigated in [32], e.g., the amplitude and phase estimation [35]. In this section, we estimate the pitch from a set of nonparametric

frequency estimates of the harmonics. We call these unconstrained frequency estimates (UFEs), which are the locations of the peaks in the spectral density estimate given the number of harmonics. Although these estimates include a coarse estimate of the fundamental frequency, we develop an optimal solution considering the harmonic signal model and noise statistics.

3.1 Single-Channel Frequency Filtering

Following the signal approximations in (B.7) and (B.13), the UFEs can be expressed as the sum of the true frequencies of the harmonics plus the frequency estimation errors $\Delta\mathbf{\Omega}_m$ due to the phase-noise vector $\Delta\mathbf{\Psi}_m$ for $m = 0$, i.e.,

$$\begin{aligned}\hat{\mathbf{\Omega}}_m &= \begin{bmatrix}\hat{\omega}_1 & \hat{\omega}_2 & \dots & \hat{\omega}_L\end{bmatrix}^T \\ &\triangleq \mathbf{\Omega} + \Delta\mathbf{\Omega}_m. \end{aligned} \tag{B.14}$$

We consider the UFEs of a ML frequency estimator as multivariate random variables with the non-zero frequencies $\lim_{N\to\infty} \mathrm{E}\{\hat{\mathbf{\Omega}}_m\} = \mathbf{\Omega}$ and the covariance matrix which is given by [33]

$$\mathbf{R}_{\Delta\mathbf{\Omega}_m} = \frac{12}{N(N^2-1)}\,\mathbf{R}_{\Delta\mathbf{\Psi}_m}. \tag{B.15}$$

This expression reveals that the variance of the frequency estimates is directly related to the corresponding narrowband power spectrum of the noise. Therefore, we can estimate the narrowband SNRs of the harmonics from statistics of the UFEs.

We estimate the fundamental frequency from the weighted sum of the UFEs by apply a real-valued linear filter, $\mathbf{h}_\omega \in \mathbb{R}^L$, such that

$$\begin{aligned}\hat{\omega}_{0,m} &= \mathbf{h}_\omega^T\,\hat{\mathbf{\Omega}}_m \\ &= \mathbf{h}_\omega^T\,\mathbf{d}_L\,\omega_0 + \mathbf{h}_\omega^T\Delta\mathbf{\Omega}_m. \end{aligned} \tag{B.16}$$

With the distortionless constraint that $\mathbf{h}_\omega^T\,\mathbf{d}_L = 1$, the variance of the estimator is given by

$$\begin{aligned}\mathrm{E}\{(\hat{\omega}_{0,m} - \omega_0)^2\} &= \mathrm{E}\{(\mathbf{h}_\omega^T\Delta\mathbf{\Omega}_m)(\Delta\mathbf{\Omega}_m^T\mathbf{h}_\omega)\} \\ &= \mathbf{h}_\omega^T\mathbf{R}_{\Delta\mathbf{\Omega}_m}\mathbf{h}_\omega. \end{aligned} \tag{B.17}$$

We design an optimal filter to minimize the variance with the distortionless constraint, i.e.,

$$\begin{aligned}\min_{\mathbf{h}_\omega} \quad & \mathbf{h}_\omega^T\mathbf{R}_{\Delta\mathbf{\Omega}_m}\mathbf{h}_\omega \\ \text{subject to} \quad & \mathbf{h}_\omega^T\,\mathbf{d}_L = 1. \end{aligned} \tag{B.18}$$

The minimum variance distortionless response (MVDR) filter is the solution that has the following form:

$$\begin{aligned}\mathbf{h}_{\omega,\text{MVDR}} &= \mathbf{R}_{\Delta\boldsymbol{\Omega}_m}^{-1}\mathbf{d}_L(\mathbf{d}_L^T\,\mathbf{R}_{\Delta\boldsymbol{\Omega}_m}^{-1}\mathbf{d}_L)^{-1} \\ &= \frac{1}{\sum_{l=1}^{L}\frac{l^2\alpha_l^2}{\Phi_m(\omega_l)}}\begin{bmatrix}\frac{\alpha_1^2}{\Phi_m(\omega_1)} & \cdots & \frac{L\alpha_L^2}{\Phi_m(\omega_L)}\end{bmatrix}^T.\end{aligned} \tag{B.19}$$

Depending on the color of the Gaussian noise, the power spectrum of the noise is different across the frequencies. In white noise, which has a uniform power spectrum, the optimal filter can be simplified in the specific case. The weights of the MVDR filter are replaced by the weights which relate to the squared amplitudes of the harmonics. The design is eventually the same as the weighted least squares (WLS) pitch estimator [7] in white noise as

$$\mathbf{h}_{\omega,\text{WLS}} = \frac{1}{\sum_{l=1}^{L} l^2\alpha_l^2}\begin{bmatrix}\alpha_1^2 & 2\alpha_2^2 & \cdots & L\alpha_L^2\end{bmatrix}^T. \tag{B.20}$$

Regarding to the noisy signal approximation (B.7), we can use the narrowband power spectrum of the UFEs for the squared amplitudes of the harmonics.

3.2 Multichannel Frequency Filtering

We can extend the single-channel pitch estimator of a single source to a solution for multichannel signals in a sequential process. Firstly, we estimate the fundamental frequency from the UFEs of M channels, individually. The estimates are

$$\begin{bmatrix}\hat{\omega}_{0,0} & \hat{\omega}_{0,1} & \cdots & \hat{\omega}_{0,M-1}\end{bmatrix}^T \triangleq \mathbf{1}_M\omega_0 + \Delta\boldsymbol{\Omega}_M, \tag{B.21}$$

where $\mathbf{1}_M$ is the all-ones column vector of length M, and $\Delta\boldsymbol{\Omega}_M$ is a vector consists of the estimation errors. We assume that the noises among the microphones are uncorrelated, the covariance matrix $\mathbf{R}_{\Delta\boldsymbol{\Omega}_M} = \mathrm{E}\{\Delta\boldsymbol{\Omega}_M\Delta\boldsymbol{\Omega}_M^T\}$ is therefore diagonal and formed by the variance of the individual estimates, i.e., $\mathrm{E}\{(\hat{\omega}_{0,m}-\omega_0)^2\} = (\mathbf{d}_L^T\,\mathbf{R}_{\Delta\boldsymbol{\Omega}_m}^{-1}\mathbf{d}_L)^{-1}$. Secondly, we apply a real-valued post-filter, $\mathbf{g}_\omega$, of length M at the individual estimates in the previous step, and design an optimal filter such that

$$\begin{aligned}\min_{\mathbf{g}_\omega} \quad & \mathbf{g}_\omega^T\mathbf{R}_{\Delta\boldsymbol{\Omega}_M}\mathbf{g}_\omega \\ \text{subject to} \quad & \mathbf{g}_\omega^T\mathbf{1}_M = 1.\end{aligned} \tag{B.22}$$

We find the optimal solution:

$$\mathbf{g}_{\omega,\text{MVDR}} = \mathbf{R}^{-1}_{\Delta\boldsymbol{\Omega}_M}\mathbf{1}_M(\mathbf{1}_M^T\,\mathbf{R}^{-1}_{\Delta\boldsymbol{\Omega}_M}\mathbf{1}_M)^{-1}. \tag{B.23}$$

The variance of the estimator is $\mathbf{g}^T_{\omega,\text{MVDR}}\mathbf{R}_{\Delta\boldsymbol{\Omega}_M}\mathbf{g}_{\omega,\text{MVDR}} = (\mathbf{1}_M^T\,\mathbf{R}^{-1}_{\Delta\boldsymbol{\Omega}_M}\mathbf{1}_M)^{-1}$, and in the case that the noise power spectrum is identical among the microphones, the variance is

$$\begin{aligned}\mathbf{g}^T_{\omega,\text{MVDR}}\mathbf{R}_{\Delta\boldsymbol{\Omega}_M}\mathbf{g}_{\omega,\text{MVDR}} &= \frac{1}{M}(\mathbf{d}_L^T\,\mathbf{R}^{-1}_{\Delta\boldsymbol{\Omega}_m}\mathbf{d}_L)^{-1} \\ &= \frac{12}{NM(N^2-1)}(\mathbf{d}_L^T\,\mathbf{R}^{-1}_{\Delta\boldsymbol{\Psi}_m}\mathbf{d}_L)^{-1}.\end{aligned} \tag{B.24}$$

4 DOA Estimation

This section presents a harmonic model-based DOA estimator based on the properties of the multichannel harmonic signals. The estimator applies the multichannel phase estimates of the given harmonics of a periodic signal in the following two-steps process.

4.1 Multichannel Phase Filtering

According to the harmonic signal model in (B.8), the phase of the lth harmonic in the $(m+1)$th microphone is

$$\psi_{l,m} = \psi_l - \omega_{l,s}m. \tag{B.25}$$

The collection of multichannel phases lies on a continuous line that originates from the first microphone as

$$\begin{aligned}\boldsymbol{\psi}_l &= \begin{bmatrix}\psi_{l,0} & \psi_{l,1} & \dots & \psi_{l,M-1}\end{bmatrix}^T \\ &= \boldsymbol{\Pi}_M\begin{bmatrix}\psi_l \\ \omega_{l,\text{s}}\end{bmatrix},\end{aligned} \tag{B.26}$$

where the matrix defined as $\boldsymbol{\Pi}_M = \begin{bmatrix}\mathbf{1}_M & (\mathbf{1}_M - \mathbf{d}_M)\end{bmatrix} \in \mathbb{R}^{M\times 2}$ is based upon the number of microphones and the linear relationship between the TDOAs of the ULA and $\mathbf{d}_M = \begin{bmatrix}1 & 2 & \dots & M\end{bmatrix}^T$. Following the noisy signal approximation in (B.13), the multichannel phase estimates $\hat{\boldsymbol{\psi}}_l$ are distorted by the multichannel phase-noise vector $\Delta\boldsymbol{\psi}_l = \begin{bmatrix}\Delta\psi_{l,1} & \Delta\psi_{l,2} & \dots & \Delta\psi_{l,M}\end{bmatrix}^T$, i.e.,

$$\hat{\boldsymbol{\psi}}_l \triangleq \boldsymbol{\psi}_l + \Delta\boldsymbol{\psi}_l. \tag{B.27}$$

The covariance matrix of the multichannel phase-noise vector of a ML phase estimator given the frequency [33] is given by

$$\mathbf{R}_{\Delta\boldsymbol{\psi}_l} = \frac{1}{2N\alpha_l^2}\,\mathrm{diag}\Big\{\Phi_0(\omega_l)\;\;\Phi_1(\omega_l)\;\;\cdots\;\;\Phi_{M-1}(\omega_l)\Big\}. \tag{B.28}$$

We estimate the parameter vector $\begin{bmatrix}\psi_l & \omega_{l,\mathrm{s}}\end{bmatrix}^T$ by applying a real-valued filter, $\mathbf{H}_{\boldsymbol{\psi}_l} \in \mathbb{R}^{M\times 2}$, such that

$$\begin{aligned}\begin{bmatrix}\hat{\psi}_l \\ \hat{\omega}_{l,\mathrm{s}}\end{bmatrix} &= \mathbf{H}_{\boldsymbol{\psi}_l}^T\hat{\boldsymbol{\psi}}_l \\ &= \mathbf{H}_{\boldsymbol{\psi}_l}^T\boldsymbol{\Pi}_M\begin{bmatrix}\psi_l \\ \omega_{l,\mathrm{s}}\end{bmatrix} + \mathbf{H}_{\boldsymbol{\psi}_l}^T\Delta\boldsymbol{\psi}_l.\end{aligned} \tag{B.29}$$

With the distortionless constraint that $\mathbf{H}_{\boldsymbol{\psi}_l}^T\boldsymbol{\Pi}_M = \mathbf{I}_{2\times 2}$, the total variance of the joint estimator is given by

$$\begin{aligned}\mathrm{E}\left\{\left\|\begin{bmatrix}\hat{\psi}_l \\ \hat{\omega}_{l,\mathrm{s}}\end{bmatrix} - \begin{bmatrix}\psi_l \\ \omega_{l,\mathrm{s}}\end{bmatrix}\right\|_2^2\right\} &= \mathrm{E}\Big\{\mathrm{tr}\big\{(\mathbf{H}_{\boldsymbol{\psi}_l}^T\Delta\boldsymbol{\psi}_l)(\Delta\boldsymbol{\psi}_l^T\mathbf{H}_{\boldsymbol{\psi}_l})\big\}\Big\} \\ &= \mathrm{tr}\big\{\mathbf{H}_{\boldsymbol{\psi}_l}^T\mathbf{R}_{\Delta\boldsymbol{\psi}_l}\mathbf{H}_{\boldsymbol{\psi}_l}\big\},\end{aligned} \tag{B.30}$$

where $\mathbf{I}_{2\times 2}$ is an identity matrix of size two-by-two, $\mathrm{tr}\{\cdot\}$ is the trace of a square matrix, and $\|\cdot\|_2$ denotes the ℓ_2-norm. A minimum variance of the parameter vector estimate is achieved by solving the following optimization problem

$$\begin{aligned}\min_{\mathbf{H}_{\boldsymbol{\psi}_l}} \quad & \mathrm{tr}\big\{\mathbf{H}_{\boldsymbol{\psi}_l}^T\mathbf{R}_{\Delta\boldsymbol{\psi}_l}\mathbf{H}_{\boldsymbol{\psi}_l}\big\} \\ \text{subject to} \quad & \mathbf{H}_{\boldsymbol{\psi}_l}^T\boldsymbol{\Pi}_M = \mathbf{I}_{2\times 2}.\end{aligned} \tag{B.31}$$

Using the method of Lagrange multipliers, we obtain

$$\mathbf{H}_{\boldsymbol{\psi}_l,\mathrm{MVDR}} = \mathbf{R}_{\Delta\boldsymbol{\psi}_l}^{-1}\boldsymbol{\Pi}_M(\,\boldsymbol{\Pi}_M^T\mathbf{R}_{\Delta\boldsymbol{\psi}_l}^{-1}\boldsymbol{\Pi}_M)^{-1}. \tag{B.32}$$

Applying the optimal solution $\mathbf{H}_{\boldsymbol{\psi}_l,\mathrm{MVDR}}$ on the multichannel phase estimates, we obtain the spatial frequency of the lth harmonic as $\hat{\omega}_{l,\mathrm{s}} = \omega_{l,\mathrm{s}} + \Delta\omega_{l,s}$. The estimation error has a normal distribution with the variance

$$\mathrm{E}\big\{\Delta\omega_{l,\mathrm{s}}^2\big\} = \Big[(\,\boldsymbol{\Pi}_M^T\mathbf{R}_{\Delta\boldsymbol{\psi}_l}^{-1}\boldsymbol{\Pi}_M)^{-1}\Big]_{2,2},$$

where $[\cdot]_{i,i}$ denotes the ith diagonal element of a square matrix. In the case of identical noise power spectrum among the microphones, the variance is

$$\mathrm{E}\big\{\Delta\omega_{l,\mathrm{s}}^2\big\} = \frac{6\,\Phi_m(\omega_l)}{\alpha_l^2 NM(M^2-1)}. \tag{B.33}$$

4.2 DOA Filtering

The DOA of each harmonic can be estimated individually from the spatial frequency estimates in the previous step such that $\hat{\theta}_l = \sin^{-1}(\hat{\omega}_{l,\mathrm{s}}/l\omega_0 f_\mathrm{s}\tau_0)$. The first-order approximation of the spatial frequency estimates with respect to individual DOA estimates is given by

$$\begin{aligned}\hat{\omega}_{l,\mathrm{s}} &= \omega_l f_\mathrm{s}\tau_0 \sin(\theta_l + \Delta\theta_l) && \text{(B.34)}\\ &\approx \omega_{l,\mathrm{s}} + \omega_l f_\mathrm{s}\tau_0 \Delta\theta_l \cos(\theta_0), && \text{(B.35)}\end{aligned}$$

where the estimation error is $\Delta\theta_l \approx \Delta\omega_{l,\mathrm{s}}/\omega_l f_s\tau_0\cos(\theta_0)$ and it has a normal distribution as well as the spatial frequency estimate. In fact, the DOAs of the harmonics are equal, i.e.,

$$\boldsymbol{\Theta} \triangleq \begin{bmatrix}\theta_1 & \theta_2 & \dots & \theta_L\end{bmatrix}^T = \mathbf{1}_L\theta_0, \tag{B.36}$$

and their individual DOA estimates are

$$\begin{aligned}\hat{\boldsymbol{\Theta}} &= \begin{bmatrix}\hat{\theta}_1 & \hat{\theta}_2 & \dots & \hat{\theta}_L\end{bmatrix}^T\\ &\triangleq \boldsymbol{\Theta} + \Delta\boldsymbol{\Theta}, && \text{(B.37)}\end{aligned}$$

where

$$\begin{aligned}\Delta\boldsymbol{\Theta} &= \begin{bmatrix}\Delta\theta_1 & \Delta\theta_2 & \dots & \Delta\theta_L\end{bmatrix}^T\\ &\approx \frac{1}{\omega_0 f_s\tau_0\cos(\theta_0)}\begin{bmatrix}\Delta\omega_{1,\mathrm{s}} & \frac{\Delta\omega_{2,\mathrm{s}}}{2} & \dots & \frac{\Delta\omega_{L,\mathrm{s}}}{L}\end{bmatrix}^T.\end{aligned}$$

The covariance matrix of the DOA estimates given the spatial frequency estimates is

$$\begin{aligned}\mathbf{R}_{\Delta\boldsymbol{\Theta}} &= \mathrm{E}\{\Delta\boldsymbol{\Theta}\,\Delta\boldsymbol{\Theta}^T\}\\ &= Q^2\,\mathrm{diag}\Big\{\mathrm{E}\{\Delta\omega_{1,\mathrm{s}}^2\} \;\; \frac{\mathrm{E}\{\Delta\omega_{2,\mathrm{s}}^2\}}{4} \;\; \cdots \;\; \frac{\mathrm{E}\{\Delta\omega_{L,\mathrm{s}}^2\}}{L^2}\Big\},\end{aligned}$$

where $Q = 1/\omega_0 f_\mathrm{s}\tau_0\cos(\theta_0)$. Substituting (B.33) into the covariance matrix results in

$$\begin{aligned}\mathbf{R}_{\Delta\boldsymbol{\Theta}} &= \frac{6Q^2}{NM(M^2-1)}\,\mathrm{diag}\Big\{\frac{\Phi_m(\omega_1)}{\alpha_1^2} \;\; \cdots \;\; \frac{\Phi_m(\omega_L)}{L^2\alpha_L^2}\Big\}\\ &= \frac{(N^2-1)Q^2}{M(M^2-1)}\,\boldsymbol{\Gamma}_L^2\mathbf{R}_{\Delta\Omega_m}, && \text{(B.38)}\end{aligned}$$

where $\boldsymbol{\Gamma}_L = \mathrm{diag}\{1 \;\; 1/2 \;\; \cdots \;\; 1/L\}$. It shows how much uncertainty there is among the harmonics due to the spatial frequency estimation error in the

previous step, and their variance relates not only to the reciprocal of narrowband SNRs, but also to the order of the harmonics. In other words, in the case of uniform narrowband SNRs, the DOA estimates of high frequencies have less variance than the DOA estimate of the fundamental frequency.

We estimate the DOA by applying the filter $\mathbf{h}_\theta \in \mathbb{R}^L$ with the distortionless constraint $\mathbf{h}_\theta^T \mathbf{1}_L = 1$ on the DOA estimates of the harmonics such that

$$\begin{aligned}\hat{\theta}_0 &= \mathbf{h}_\theta^T \hat{\boldsymbol{\Theta}} \\ &= \mathbf{h}_\theta^T \mathbf{1}_L \theta_0 + \mathbf{h}_\theta^T \Delta\boldsymbol{\Theta}.\end{aligned} \tag{B.39}$$

The variance of the result is given by

$$\begin{aligned}\mathrm{E}\{(\hat{\theta}_0 - \theta_0)^2\} &= \mathrm{E}\{(\mathbf{h}_\theta^T \Delta\boldsymbol{\Theta})(\Delta\boldsymbol{\Theta}^T \mathbf{h}_\theta)\} \\ &= \mathbf{h}_\theta^T \mathbf{R}_{\Delta\boldsymbol{\Theta}} \mathbf{h}_\theta.\end{aligned} \tag{B.40}$$

Minimizing the variance subjected to the distortionless constraint, we design the MVDR filter then

$$\mathbf{h}_{\theta,\mathrm{MVDR}} = \mathbf{R}_{\Delta\boldsymbol{\Theta}}^{-1} \mathbf{1}_L \left(\mathbf{1}_L^T \mathbf{R}_{\Delta\boldsymbol{\Theta}}^{-1} \mathbf{1}_L\right)^{-1}. \tag{B.41}$$

Substituting (B.38) into (B.41), leads to the interesting expression

$$\mathbf{h}_{\theta,\mathrm{MVDR}} = \frac{1}{\sum_{l=1}^{L} \frac{l^2 \alpha_l^2}{\Phi_m(\omega_l)}} \left[\frac{\alpha_1^2}{\Phi_m(\omega_1)} \quad \dots \quad \frac{L^2 \alpha_L^2}{\Phi_m(\omega_L)} \right]^T. \tag{B.42}$$

In white Gaussian noise, we reach interestingly to the WLS DOA estimator [16] from the designed optimal filter such that

$$\mathbf{h}_{\theta,\mathrm{WLS}} = \frac{1}{\sum_{l=1}^{L} l^2 \alpha_l^2} \left[\alpha_1^2 \quad 4\alpha_2^2 \quad \dots \quad L^2 \alpha_L^2 \right]^T. \tag{B.43}$$

5 Joint DOA and Pitch Estimation

In the previous two sections, we estimated the pitch and the DOA of a periodic signal by minimizing the phase-noise with respect to either the temporal frequency or the spatial frequency, holding the other one fixed. These approaches are limited to the case when the harmonics are well separated and the narrowband SNRs of the harmonics are high. In such a situation, spatio-temporal signal processing characterizes both the temporal and the spatial frequencies at the same time, however, computationally are more complex.

Algorithm 1 Joint DOA and pitch estimation for K sources.

1: Estimate the two-dimensional spectral density $J(\theta, \omega)$.
2: Estimate initial DOAs of the known K sources.
3: Estimate the model orders $\{L_k\}_{k=1}^{K}$ from $J(\hat{\theta}_{\text{int},k}, \omega)$.
4: Localize $\{L_k\}_{k=1}^{K}$ peaks in $\{\hat{\theta}_{\text{int},k}\}_{k=1}^{K} \pm \eta$ to find $\{\hat{\mathbf{\Lambda}}_k\}_{k=1}^{K}$, where η is the DOA neighboring bound.
5: Estimate the DOA and the pitch of K sources separately: $\begin{bmatrix} \hat{\omega}_{0,k} & \hat{\theta}_{0,k} \end{bmatrix}^T = \mathbf{H}_{(\omega_k,\theta_k)}^T \hat{\mathbf{\Lambda}}_k$.

Therefore, we consider the properties of the spatio-temporal signals, and localize the peaks in a two-dimensional (2D) spectral density estimate to find the frequencies and DOAs of the harmonics such that

$$\hat{\mathbf{\Lambda}} = \begin{bmatrix} \hat{\mathbf{\Omega}} \\ \hat{\mathbf{\Theta}} \end{bmatrix} \triangleq \mathbf{\Lambda} + \Delta\mathbf{\Lambda}, \tag{B.44}$$

where $\mathbf{\Lambda} = \begin{bmatrix} \mathbf{d}_L^T \omega_0 & \mathbf{1}_L^T \theta_0 \end{bmatrix}^T$ is constituted by the true frequency and DOA values, and $\Delta\mathbf{\Lambda} = \begin{bmatrix} \Delta\mathbf{\Omega}^T & \Delta\mathbf{\Theta}^T \end{bmatrix}^T$ are the errors caused by the phase-noise. These errors have a multivariate normal distribution with zero means and the covariance matrix given by (it is proved in the next section)

$$\mathbf{R}_{\Delta\mathbf{\Lambda}} = \begin{bmatrix} \mathbf{R}_{\Delta\mathbf{\Omega}} & \frac{\tan(\theta_0)}{\omega_0}\mathbf{\Gamma}_L \mathbf{R}_{\Delta\mathbf{\Omega}} \\ \frac{\tan(\theta_0)}{\omega_0}\mathbf{\Gamma}_L \mathbf{R}_{\Delta\mathbf{\Omega}} & \mathbf{R}_{\Delta\mathbf{\Theta}} + \frac{\tan^2(\theta_0)}{\omega_0^2}\mathbf{\Gamma}_L^2 \mathbf{R}_{\Delta\mathbf{\Omega}} \end{bmatrix}, \tag{B.45}$$

where $\mathbf{R}_{\Delta\mathbf{\Omega}} = \mathbf{R}_{\Delta\mathbf{\Omega}_m}/M$ when the power spectrum of noise is identical among the microphones.

We Apply a real-valued filter, $\mathbf{H}_{(\omega,\theta)} \in \mathbb{R}^{2L\times 2}$, on the frequency and DOA estimates of the harmonic, $\hat{\mathbf{\Lambda}}$, as follows:

$$\begin{aligned} \begin{bmatrix} \hat{\omega}_0 \\ \hat{\theta}_0 \end{bmatrix} &= \mathbf{H}_{(\omega,\theta)}^T \hat{\mathbf{\Lambda}} \\ &= \mathbf{H}_{(\omega,\theta)}^T \begin{bmatrix} \mathbf{d}_L \omega_0 \\ \mathbf{1}_L \theta_0 \end{bmatrix} + \mathbf{H}_{(\omega,\theta)}^T \Delta\mathbf{\Lambda}. \end{aligned} \tag{B.46}$$

Using the distortionless constraints given in the previous two sections, the total variance of the joint estimates is

$$\begin{aligned} \mathrm{E}\left\{ \left\| \begin{bmatrix} \hat{\omega}_0 \\ \hat{\theta}_0 \end{bmatrix} - \begin{bmatrix} \omega_0 \\ \theta_0 \end{bmatrix} \right\|_2^2 \right\} &= \mathrm{E}\Big\{ \mathrm{tr}\big\{ (\mathbf{H}_{(\omega,\theta)}^T \Delta\mathbf{\Lambda})(\Delta\mathbf{\Lambda}^T \mathbf{H}_{(\omega,\theta)}) \big\} \Big\} \\ &= \mathrm{tr}\Big\{ \mathbf{H}_{(\omega,\theta)}^T \mathbf{R}_{\Delta\mathbf{\Lambda}} \mathbf{H}_{(\omega,\theta)} \Big\}. \end{aligned} \tag{B.47}$$

Since the true frequencies and DOAs are independent, we define $\mathbf{H}_{(\omega,\theta)} \triangleq \begin{bmatrix} \mathbf{f}_\omega & \mathbf{0} \\ \mathbf{0} & \mathbf{f}_\theta \end{bmatrix}$, where $\mathbf{f}_\omega$ and $\mathbf{f}_\theta$ are real-valued vectors holding the distortionless constraints, and $\mathbf{0}$ is the zero vector of length L. An optimal filter is designed to approach the minimum variance of the joint estimates as

$$\min_{\mathbf{H}_{(\omega,\theta)}} \operatorname{tr}\Big\{\mathbf{H}_{(\omega,\theta)}^T \mathbf{R}_{\Delta\mathbf{\Lambda}} \mathbf{H}_{(\omega,\theta)}\Big\} = \mathbf{f}_\omega^T \mathbf{R}_{\Delta\mathbf{\Omega}} \mathbf{f}_\omega + \mathbf{f}_\theta^T \Big(\mathbf{R}_{\Delta\mathbf{\Theta}} + \frac{\tan^2(\theta_0)}{\omega_0^2} \mathbf{\Gamma}_L^2 \mathbf{R}_{\Delta\mathbf{\Omega}}\Big) \mathbf{f}_\theta \tag{B.48}$$

$$\text{subject to } \mathbf{f}_\omega^T \mathbf{d}_L = 1, \text{ and } \mathbf{f}_\theta^T \mathbf{1}_L = 1.$$

The MVDR filter is designed as

$$\mathbf{H}_{(\omega,\theta),\text{MVDR}} = \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \end{bmatrix}, \tag{B.49}$$

where

$$\mathbf{A} = \mathbf{R}_{\Delta\mathbf{\Omega}}^{-1} \mathbf{d}_L (\mathbf{d}_L^T \mathbf{R}_{\Delta\mathbf{\Omega}}^{-1} \mathbf{d}_L)^{-1},$$
$$\mathbf{B} = \Big(\mathbf{R}_{\Delta\mathbf{\Theta}} + \frac{\tan^2(\theta_0)}{\omega_0^2} \mathbf{\Gamma}_L^2 \mathbf{R}_{\Delta\mathbf{\Omega}}\Big)^{-1} \mathbf{1}_L \Big[\mathbf{1}_L^T \Big(\mathbf{R}_{\Delta\mathbf{\Theta}} + \frac{\tan^2(\theta_0)}{\omega_0^2} \mathbf{\Gamma}_L^2 \mathbf{R}_{\Delta\mathbf{\Omega}}\Big)^{-1} \mathbf{1}_L\Big]^{-1}.$$

In white noise, the WLS estimator results as

$$\mathbf{H}_{(\omega,\theta),\text{WLS}} = \frac{1}{\sum_{l=1}^{L} (l\alpha_l)^2} \begin{bmatrix} \begin{bmatrix} \alpha_1^2 & 2\alpha_2^2 & \dots & L\alpha_L^2 \end{bmatrix}^T & & \mathbf{0} & \\ & \mathbf{0} & & \begin{bmatrix} \alpha_1^2 & 4\alpha_2^2 & \dots & L^2\alpha_L^2 \end{bmatrix}^T \end{bmatrix} \tag{B.50}$$

The joint and the multichannel pitch estimators have the same variance, i.e., $(\mathbf{d}_L^T \mathbf{R}_{\Delta\mathbf{\Omega}}^{-1} \mathbf{d}_L)^{-1}$. However, the variance of the DOA estimate in the joint estimator is greater than (or equal to) the DOA estimate given the frequencies of the harmonics, i.e.,

$$\Big[\mathbf{1}_L^T \Big(\mathbf{R}_{\Delta\mathbf{\Theta}} + \frac{\tan^2(\theta_0)}{\omega_0^2} \mathbf{\Gamma}_L^2 \mathbf{R}_{\Delta\mathbf{\Omega}}\Big)^{-1} \mathbf{1}_L\Big]^{-1} \geq (\mathbf{1}_L^T \mathbf{R}_{\Delta\mathbf{\Theta}}^{-1} \mathbf{1}_L)^{-1}. \tag{B.51}$$

This interesting statement is derived as well as the minimum variance of the DOA estimate in the next section.

5.1 Multiple Sources Estimates

We can potentially separate multiple signal sources which are located at different positions at the same time. Therefore, the joint DOA and pitch estimator can be used in multiple source scenarios. The general algorithm to

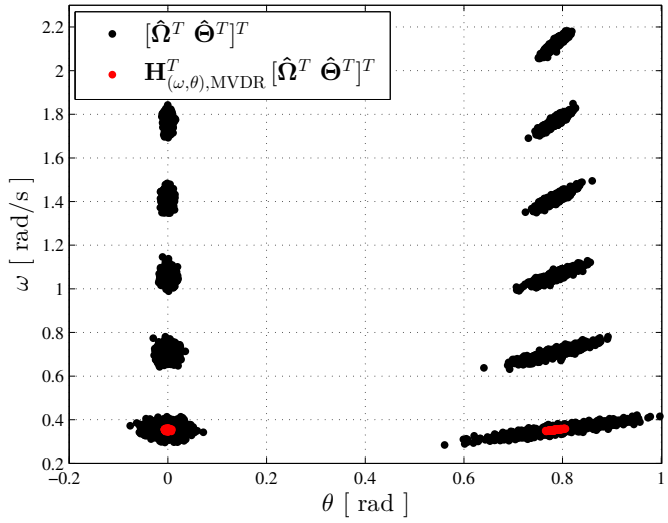


Fig. B.1: Scatter plot of (black) DOAs and frequencies of the harmonics of two sources, and (red) joint DOA and pitch estimates of the sources.

estimate the parameters of K sources is outlined in Alg. 1. Clearly, we require initial DOAs of the signal sources to estimate the number of harmonics (model order) L_k and the join parameter estimates $\hat{\mathbf{\Lambda}}_k$ for $k = 1, 2, \ldots, K$, respectively in steps 3 and 4. Therefore, we apply a broadband nonparametric DOA estimator using the 2D spectral density $J(\theta, \omega)$ for $\omega \in [0, \pi]$ and $\theta \in [-\pi/2, \pi/2]$ such that

$$\{\hat{\theta}_{\mathrm{int},k}\}_{k=1}^{K} = \arg \max_{\{\theta_k\}_{k=1}^{K}} \int J(\theta_k, \omega) d\omega. \tag{B.52}$$

To estimate the model order of each source, we incorporate the Bayesian information criterion (BIC) [36] (see also [37]) with the 2D spectral density estimate [38]. As it turns out, we localize L_k peaks in a range of DOAs to estimate $\hat{\mathbf{\Lambda}}_k$. We design the filters $\mathbf{H}_{(\omega_k,\theta_k)}$ for each source separately. As an example, we apply the covariance matrix in (B.45) to model two signals that have $L_1 = 5$ and $L_2 = 6$ harmonics at the angles $\theta_{0,1} = 0$ and $\theta_{0,2} = \pi/4$ radians, respectively, in white noise. Fig. B.1 illustrates the scatter plot of Monte-Carlo frequency and DOA estimates of the harmonics as well as the joint estimates using the proposed MVDR method.

6 Performance Analysis

The most commonly used benchmark for assessing the accuracy of a deterministic parameter estimate is the lowest possible mean squared error (MSE), which is equal to the minimum variance of an unbiased estimate and called the Cramér-Rao lower bound (CRLB). The CRLB of DOA and pitch estimates in white noise had been formulated in [26], which also showed that the asymptotic bounds approach the exact bounds for large spatial and temporal samples. In this section, we derive the asymptotic CRLB of those parameters in general Gaussian noise by relying on the signal approximations in (B.7) and (B.13).

6.1 Single-Channel Pitch Estimate

We write the real-valued arguments of the harmonics of a noisy signal as $\Xi_0(n) \triangleq \mathbf{d}_L \omega_0 n + \boldsymbol{\Psi} + \Delta\boldsymbol{\Psi}_0$, where $\boldsymbol{\Psi} = \begin{bmatrix} \psi_1 & \psi_2 & \dots & \psi_L \end{bmatrix}^T$. The likelihood function of these arguments is modeled as a multivariate normal distribution with non-zero means as a linear combination of the auxiliary parameter vector $\boldsymbol{\xi} = \begin{bmatrix} \omega_0 & \boldsymbol{\Psi}^T \end{bmatrix}^T$ and the covariance matrix $\mathbf{R}_{\Delta\boldsymbol{\Psi}_0}$, i.e.,

$$\boldsymbol{p}(\Xi_0(n), \boldsymbol{\xi}) \sim \mathcal{N}(\mathbf{d}_L \omega_0 n + \boldsymbol{\Psi}, \mathbf{R}_{\Delta\boldsymbol{\Psi}_0}). \tag{B.53}$$

For N independent samples $\Xi_0 \triangleq \{\Xi_0(n)\}_{n=0}^{n=N-1}$ with the likelihood function of $\boldsymbol{p}(\Xi_0, \boldsymbol{\xi})$, the Fisher information matrix (FIM), $\mathbf{I}(\boldsymbol{\xi})$ of the given parameter vector is

$$\begin{aligned} \mathbf{I}(\boldsymbol{\xi}) &= \begin{bmatrix} -\mathrm{E}\{\frac{\partial^2 \ln \boldsymbol{p}(\Xi_0,\boldsymbol{\xi})}{\partial \omega_0^2}\} & -\mathrm{E}\{\frac{\partial^2 \ln \boldsymbol{p}(\Xi_0,\boldsymbol{\xi})}{\partial \omega_0 \partial \boldsymbol{\Psi}^T}\} \\ -\mathrm{E}\{\frac{\partial^2 \ln \boldsymbol{p}(\Xi_0,\boldsymbol{\xi})}{\partial \boldsymbol{\Psi} \partial \omega_0}\} & -\mathrm{E}\{\frac{\partial^2 \ln \boldsymbol{p}(\Xi_0,\boldsymbol{\xi})}{\partial \boldsymbol{\Psi} \partial \boldsymbol{\Psi}^T}\} \end{bmatrix} \\ &= \begin{bmatrix} \frac{N(N-1)(2N-1)}{6} \mathbf{d}_L^T \mathbf{R}_{\Delta\boldsymbol{\Psi}_0}^{-1} \mathbf{d}_L & \frac{N(N-1)}{2} \mathbf{d}_L^T \mathbf{R}_{\Delta\boldsymbol{\Psi}_0}^{-1} \\ \frac{N(N-1)}{2} \mathbf{R}_{\Delta\boldsymbol{\Psi}_0}^{-1} \mathbf{d}_L & N \mathbf{R}_{\Delta\boldsymbol{\Psi}_0}^{-1} \end{bmatrix}. \end{aligned} \tag{B.54}$$

The CRLB of the ith unbiased parameter estimate, i.e., $\mathrm{E}\{\hat{\boldsymbol{\xi}}_i\} = \boldsymbol{\xi}_i$, is obtained from the inverse of the FIM such that $\mathrm{CRLB}(\boldsymbol{\xi}_i) = [\mathbf{I}(\boldsymbol{\xi})^{-1}]_{i,i}$. With the Woodbury's identity of the matrix inversion, we get the asymptotic CRLB of the fundamental frequency and the phases of the harmonics, respectively, as

$$\mathrm{CRLB}(\omega_0) = \frac{12}{N(N^2-1)} (\mathbf{d}_L^T \mathbf{R}_{\Delta\boldsymbol{\Psi}_0}^{-1} \mathbf{d}_L)^{-1}, \tag{B.55}$$

$$\mathrm{CRLB}(\psi_l) = \left[\frac{3(N-1)}{N(N+1)} \mathbf{d}_L (\mathbf{d}_L^T \mathbf{R}_{\Delta\boldsymbol{\Psi}_0}^{-1} \mathbf{d}_L)^{-1} \mathbf{d}_L^T + \frac{1}{N} \mathbf{R}_{\Delta\boldsymbol{\Psi}_0} \right]_{l,l}. \tag{B.56}$$

When the fundamental frequency is known, the FIM of the parameter vector $\boldsymbol{\xi} = \boldsymbol{\Psi}$ is $\mathbf{I}(\boldsymbol{\xi}) = N\mathbf{R}_{\Delta\boldsymbol{\Psi}_0}^{-1}$. Hence, the CRLB of the phase estimate given the frequency is

$$\mathrm{CRLB}(\psi_l|\omega_0) = \left[\frac{1}{N}\mathbf{R}_{\Delta\boldsymbol{\Psi}_0}\right]_{l,l} = \frac{\Phi_0(\omega_l)}{2N\alpha_l^2}. \tag{B.57}$$

6.2 Multichannel DOA and Pitch Estimates

We write the real-valued arguments of the harmonics in (B.13) as $\boldsymbol{\Xi}_m(n) \triangleq \mathbf{d}_L\omega_0 n - \mathbf{d}_L\omega_0 f_s\tau_0\sin(\theta_0)m + \boldsymbol{\Psi} + \Delta\boldsymbol{\Psi}_m$ and the corresponding likelihood function with respect to the auxiliary parameter vector $\boldsymbol{\xi} = \begin{bmatrix}\omega_0 & \theta_0 & \boldsymbol{\Psi}^T\end{bmatrix}^T$ as

$$\boldsymbol{p}(\boldsymbol{\Xi}_m(n), \boldsymbol{\xi}) \sim \mathcal{N}(\mathbf{d}_L\omega_0 n - \mathbf{d}_L\omega_0 f_s\tau_0\sin(\theta_0)m + \boldsymbol{\Psi}, \mathbf{R}_{\Delta\boldsymbol{\Psi}_m}).$$

We have the FIM of the real-valued arguments of $N \times M$ independent samples $\boldsymbol{\Xi} \triangleq \left\{\left\{\boldsymbol{\Xi}_m(n)\right\}_{n=0}^{n=N-1}\right\}_{m=0}^{m=M-1}$ with the likelihood function $\boldsymbol{p}(\boldsymbol{\Xi}, \boldsymbol{\xi})$ as

$$\mathbf{I}(\boldsymbol{\xi}) = \begin{bmatrix} -\mathrm{E}\{\frac{\partial^2 \ln \boldsymbol{p}(\boldsymbol{\Xi},\boldsymbol{\xi})}{\partial\omega_0^2}\} & -\mathrm{E}\{\frac{\partial^2 \ln \boldsymbol{p}(\boldsymbol{\Xi},\boldsymbol{\xi})}{\partial\omega_0\partial\theta_0}\} & -\mathrm{E}\{\frac{\partial^2 \ln \boldsymbol{p}(\boldsymbol{\Xi},\boldsymbol{\xi})}{\partial\omega_0\partial\boldsymbol{\Psi}^T}\} \\ -\mathrm{E}\{\frac{\partial^2 \ln \boldsymbol{p}(\boldsymbol{\Xi},\boldsymbol{\xi})}{\partial\theta_0\partial\omega_0}\} & -\mathrm{E}\{\frac{\partial^2 \ln \boldsymbol{p}(\boldsymbol{\Xi},\boldsymbol{\xi})}{\partial\theta_0^2}\} & -\mathrm{E}\{\frac{\partial^2 \ln \boldsymbol{p}(\boldsymbol{\Xi},\boldsymbol{\xi})}{\partial\theta_0\partial\boldsymbol{\Psi}^T}\} \\ -\mathrm{E}\{\frac{\partial^2 \ln \boldsymbol{p}(\boldsymbol{\Xi},\boldsymbol{\xi})}{\partial\boldsymbol{\Psi}\partial\omega_0}\} & -\mathrm{E}\{\frac{\partial^2 \ln \boldsymbol{p}(\boldsymbol{\Xi},\boldsymbol{\xi})}{\partial\boldsymbol{\Psi}\partial\theta_0}\} & -\mathrm{E}\{\frac{\partial^2 \ln \boldsymbol{p}(\boldsymbol{\Xi},\boldsymbol{\xi})}{\partial\boldsymbol{\Psi}\partial\boldsymbol{\Psi}^T}\} \end{bmatrix}. \tag{B.58}$$

Assuming identical covariance matrices among the microphones, the asymptotic CRLB of pitch and DOA estimates, respectively, are

$$\mathrm{CRLB}(\omega_0) = \frac{12}{NM(N^2-1)}(\mathbf{d}_L^T\mathbf{R}_{\Delta\boldsymbol{\Psi}_m}^{-1}\mathbf{d}_L)^{-1}, \tag{B.59}$$

$$\mathrm{CRLB}(\theta_0) = \Big[\frac{12}{NM(M^2-1)(\omega_0 f_s\tau_0\cos(\theta_0))^2} + \frac{12\tan^2(\theta_0)}{NM(N^2-1)\omega_0^2}\Big](\mathbf{d}_L^T\mathbf{R}_{\Delta\boldsymbol{\Psi}_m}^{-1}\mathbf{d}_L)^{-1}. \tag{B.60}$$

When the fundamental frequency is known, we derive the CRLB of the DOA estimate given the frequency through the FIM of the parameter vector $\boldsymbol{\xi} = \begin{bmatrix}\theta_0 & \boldsymbol{\Psi}^T\end{bmatrix}^T$, which yields

$$\mathrm{CRLB}(\theta_0|\omega_0) = \frac{12\,(\mathbf{d}_L^T\mathbf{R}_{\Delta\boldsymbol{\Psi}_m}^{-1}\mathbf{d}_L)^{-1}}{NM(M^2-1)(\omega_0 f_s\tau_0\cos(\theta_0))^2}. \tag{B.61}$$

The CRLBs of the joint estimates confirm the results of [26] in white Gaussian noise. In particular, the bounds depend not only on the number of the

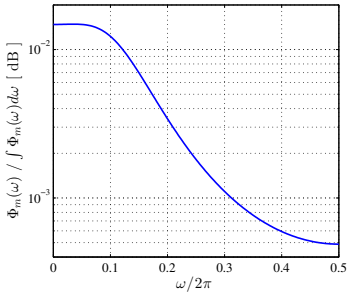


Fig. B.2: Normalized power spectrum of the colored noise.

samples and noise statistics, but also on the parameters of the signal. For example, a large number of harmonics yield low CRLBs of pitch and DOA estimates, and the lowest bound of a DOA estimate results at zero angle. Furthermore, the DOA estimate given the fundamental frequency, has a lower bound than the DOA estimate in the joint estimates, i.e., $\mathrm{CRLB}(\theta_0|\omega_0) \leq \mathrm{CRLB}(\theta_0)$.

Following the proposed joint estimator, the real-valued arguments of the harmonics are represented as $\Xi_m(n) \triangleq \mathbf{\Omega}\, n - \mathbf{\Omega} \odot \sin(\mathbf{\Theta}) f_s \tau_0 m + \mathbf{\Psi} + \Delta\mathbf{\Psi}_m$, where $\odot$ denotes the element wise product of two vectors, and $\sin(\mathbf{\Theta})$ is the vector includes the sine transform of the DOAs of the harmonics. We can then derive the covariance matrix (B.45) through the FIM of the likelihood function $p(\Xi, \boldsymbol{\xi})$ with respect to the parameter vector $\boldsymbol{\xi} = \begin{bmatrix} \mathbf{\Omega}^T & \mathbf{\Theta}^T & \mathbf{\Psi}^T \end{bmatrix}^T$.

We now proceed to evaluate the proposed estimators. First we conduct 500 Monte-Carlo simulations for the parameter setting of synthetic signals in the presence of colored noise, and then compare the mean squared error (MSE) of the results with the given CRLBs. We also conduct experiments using real audio signals. Finally, we analyze the computation complexity of the estimators.

In simulations, we generate the colored noise $v_m(n)$ by passing a complex white noise $w(n)$ with zero mean and variance σ^2 through a filter with impulse response $g(n)$. The corresponding output signal and the power spectrum are, respectively,

$$v_m(n) = \sum_{k=-\infty}^{\infty} g(k) w(n-k), \tag{B.62}$$

$$\Phi_m(\omega) = \Big| \sum_{k=-\infty}^{\infty} v_m(n) e^{-j\omega k} \Big|^2 = \sigma^2 \left| G(\omega) \right|^2, \tag{B.63}$$

where $G(\omega)$ is the Fourier transform of $g(n)$. For $G(\omega) = 1/(1 + ae^{-j\omega} + be^{-j2\omega})$, with $a = -0.9$, $b = 0.3$, and $\sigma^2 = 1$, the normalized power spectrum of the resulting signal is shown in Fig. B.2. Most energy of the colored noise resides in low frequencies and in the range of the fundamental frequency of

real audio signals. We have generated such the colored noise to mimic some real environmental noise, e.g., wind noise [19].

We estimate the covariance matrices of the parameter estimates once per frame using B earlier estimates. For example, the covariance matrix of the UFEs is calculated at the time instance n as

$$\hat{\mathbf{R}}_{\Delta\boldsymbol{\Omega}_m} = \frac{1}{B}\sum_{b=0}^{B-1} \Delta\boldsymbol{\Omega}_m(n+b)\Delta\boldsymbol{\Omega}_m^T(n+b), \tag{B.64}$$

where $\Delta\boldsymbol{\Omega}_m(n) = \hat{\boldsymbol{\Omega}}_m(n) - \hat{\boldsymbol{\mu}}_{\boldsymbol{\Omega}_m}(n)$, and $\boldsymbol{\mu}_{\boldsymbol{\Omega}_m}(n) = \mathrm{E}\{\hat{\boldsymbol{\Omega}}_m(n)\}$. We can estimate the statistical expectation using the time average of the earlier stationary parameters, however, the long-term properties of speech signals are not stationary. Therefore, we apply an exponential moving average with a forgetting factor $0 < \lambda < 1$ to update the time-varying statistical expectation of the UFEs, i.e.,

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{\Omega}_m}(n) = \lambda\,\hat{\boldsymbol{\Omega}}_m(n) + (1-\lambda)\,\hat{\boldsymbol{\mu}}_{\boldsymbol{\Omega}_m}(n-1). \tag{B.65}$$

In the estimation of the unconstrained frequencies of the harmonics, we may miss some harmonics at low narrowband SNRs due to an incorrect model order estimate. This practical problem causes a mismatch between the UFEs and the true harmonics. Though this objective itself establishes another problem, we propose a heuristic solution to estimate $\mathbf{d}_L$ of the UFEs. In Alg. 2, we compare an expected fundamental frequency with the minimum difference between the expected UFEs. Furthermore, in the covariance matrix estimation of the UFEs, the problem is twofold: short-term (ST) and long-term (LT) harmonics mismatch. For example, Figs. B.3-a and B.3-b show the case that the second harmonic of the UFEs is missed over time. The short-term mismatch causes a bias error in the covariance matrix estimate. In the long-term mismatch, we can estimate the covariance matrix successfully.

Algorithm 2 Estimate $\mathbf{d}_L$ of the UFEs

Inputs: $\hat{\boldsymbol{\Omega}}_m(n)$ and $\hat{\boldsymbol{\mu}}_{\boldsymbol{\Omega}_m}(n)$

1: $\gamma = \min\left\{\left|[\hat{\boldsymbol{\mu}}_{\boldsymbol{\Omega}_m}(n)]_h - [\hat{\boldsymbol{\mu}}_{\boldsymbol{\Omega}_m}(n)]_k\right|\right\}$ for $h \neq k$
2: **if** $\left|[\hat{\boldsymbol{\mu}}_{\boldsymbol{\Omega}_m}(n)]_1 - \gamma\right| < \epsilon$ **then**
3: $\quad \mathbf{d}_L = \lfloor \hat{\boldsymbol{\Omega}}_m(n) / [\hat{\boldsymbol{\mu}}_{\boldsymbol{\Omega}_m}(n)]_1 \rceil$
4: **else**
5: $\quad \mathbf{d}_L = \lfloor \hat{\boldsymbol{\Omega}}_m(n) / \gamma \rceil$, where $\lfloor\,\cdot\,\rceil$ means the nearest integer.
6: **end if**

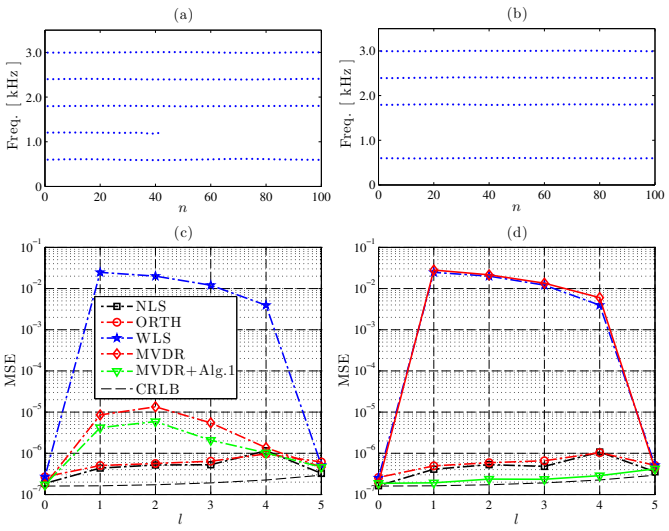


Fig. B.3: (top) Examples of missing the second harmonic over time in (a) short-term (ST) and (b) long-term (LT). (bottom) MSE of the pitch estimates for (c) the ST and (d) the LT situations plotted in dashed and solid lines, respectively, as a function of the missed harmonic.

6.3 Synthetic Signal Analysis

We generate synthetic signals for simulations using the signal model in (B.8). The signals have $L = 5$ complex sinusoids with $\omega_0 = 0.15\pi$, identical amplitudes with uniformly distributed random phases, and in the direction $\theta_0 = \pi/6$ radians. We apply $M = 5$ omnidirectional microphones for a uniform linear array (ULA) with the fixed distance between adjacent microphones, $\delta = 0.04$ m, where $c = 343.2$ m/s, and $f_s = 8.0$ kHz. All signals are corrupted with the colored Gaussian noise. The unconstrained frequencies are estimated using the multiple signal classification (MUSIC) method [39] given the number of harmonics, where the model order can be estimated in practice using a method in [37]. We estimate the amplitudes of the UFEs using the least squares as $\begin{bmatrix}\hat{\alpha}_1 & \hat{\alpha}_2 & \dots & \hat{\alpha}_L\end{bmatrix}^T = \left|(\mathbf{Z}^H\mathbf{Z})^{-1}\mathbf{Z}^H\mathbf{y}(n)\right|$, where $\mathbf{Z} = \begin{bmatrix}\mathbf{z}_t(\hat{\omega}_1) & \mathbf{z}_t(\hat{\omega}_2) & \dots & \mathbf{z}_t(\hat{\omega}_L)\end{bmatrix}$, and $\mathbf{z}_t(\hat{\omega}_l) = \begin{bmatrix}1 & e^{j\hat{\omega}_l} & \dots & e^{j(N-1)\hat{\omega}_l}\end{bmatrix}^T$. The number of samples and the broadband frequency grids are $N = 60$ and $F = 65{,}536$, respectively. We design the proposed MVDR filters using the statistics of $B = 100$ estimates, and assign them on the last estimates.

In the first experiment, we evaluate the single-channel pitch estimates of the synthetic signal versus the narrowband SNR of the fundamental frequency in Fig. B.4-a. The MSE of the proposed method is shown with the competing statistical efficient methods: the nonlinear least squares (NLS) [5], the subspace orthogonality based method (ORTH) [5, 6], and the WLS [7]. We

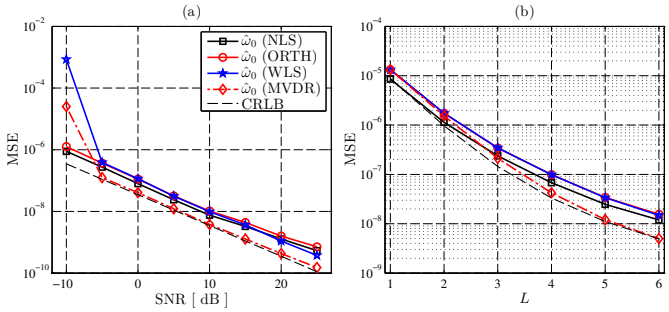


Fig. B.4: MSE of pitch estimates: (a) versus the narrowband SNR^l, $l = 1$, and (b) the number of the harmonics in $\text{SNR}^l = 5$ dB.

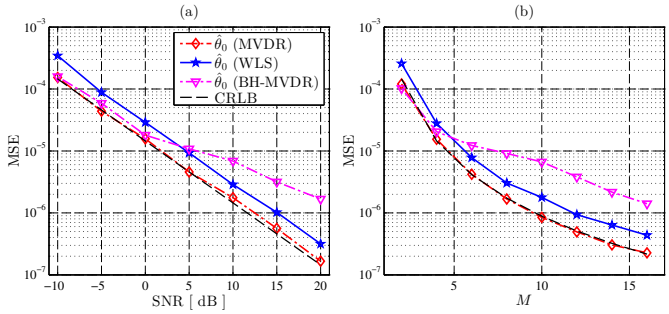


Fig. B.5: MSE of DOA estimates versus (a) the narrowband SNR^l, $l = 1$, and (b) the number of microphones in $\text{SNR}^l = 0$ dB.

can see although the NLS method has the lower MSE than the ORTH and the WLS methods, below an SNR of 15 dB, the methods have a bias error. The MSE of the proposed method is lower than the other methods, and comes close to the CRLB in the narrowband SNRs greater than or equal to -5 dB. Next, we investigate about the effect of the number of harmonics on the performance. In 5 dB SNR, Fig. B.4-b shows that the WLS and the ORTH have the same results, and however the MUSIC method has a bias error (this can be seen in a single sinusoid, i.e., $L = 1$), the proposed method compensates this bias error in large model orders, $L \geq 5$, and has a lower MSE than the NLS in $L \geq 3$.

Next, we examine the harmonics mismatching problem by conducting experiments using the synthetic signal in white noise and 5 dB narrowband SNRs at the harmonics. We simulate the ST and LT situations by changing the amplitude of $l$the harmonic to zero, where $l = 0$ means all amplitudes are not zero. We apply Alg. 2 to the MVDR filters with $\epsilon = 100 \times (2\pi/f_s)$ rad/sample. Figs. B.3-c and B.3-d show the MSE of pitch estimates in the ST and the LT situations, respectively. As can be seen, the MVDR pitch estimator

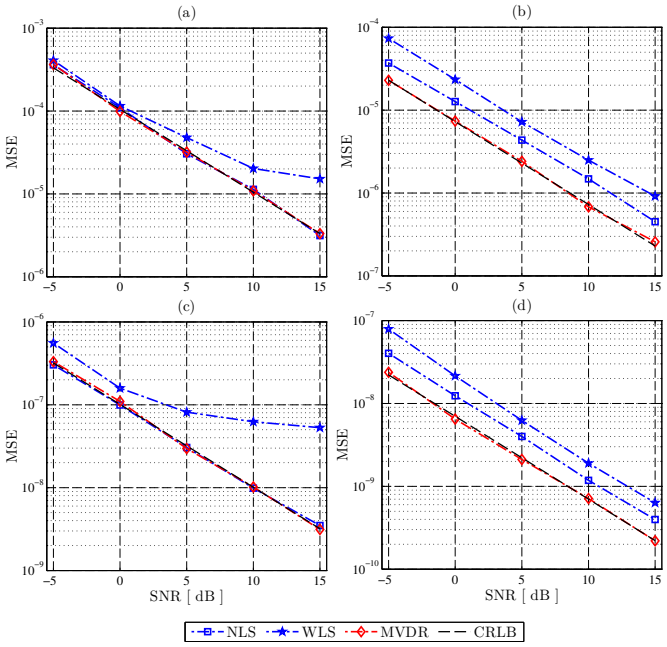


Fig. B.6: MSE of joint (a-b) DOA and (c-d) pitch estimates versus the narrowband SNR^l, $l = 1$, of (a-c) white noise and (b-d) colored noise.

has a larger MSE than the NLS and the ORTH methods in the ST situation. However, the MVDR pitch estimator that uses the Alg. 2 nearly reaches the CRLB and outperforms the competing methods in the LT situation. Moreover, these experiments confirm the importance of an accurate estimate of $\mathbf{d}_L$ in the proposed pitch estimator.

In the next experiments we explore the performance of the proposed two-step MVDR DOA estimator, in the colored noise, compared with the harmonic mode-based DOA estimators: the WLS method [16], and the estimation method using the broadband MVDR beamformer (BH-MVDR) [23]. In the BH-MVDR method, the DOA is estimated by maximizing the integrated two-dimensional (2D) spectral density estimate, of noisy signals, over the given frequencies of the harmonics. In the proposed DOA estimator, we estimate the multichannel phases of the harmonics from the arguments of the complex amplitude estimates. We apply the NLS complex amplitude estimator [5] such that $(\mathbf{Z}^H\mathbf{Z})^{-1}\mathbf{Z}^H\mathbf{y}_m(n)$, where $\mathbf{Z} = \begin{bmatrix}\mathbf{z}_t(\omega_0) & \mathbf{z}_t(2\omega_0) & \dots & \mathbf{z}_t(L\omega_0)\end{bmatrix}$. We unwrap the phase estimates using the algorithm in [40]. Fig. B.5 shows the MSE of the DOA estimate versus different narrowband SNR of the fundamental frequency, and also versus different number of microphones. The results indicate that the WLS estimator has the higher MSE compared with

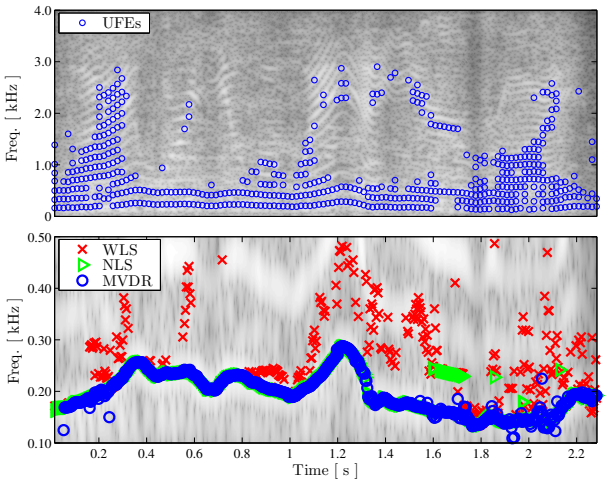


Fig. B.7: (top) Spectrogram and UFEs of a speech signal drowned in car noise. (bottom) Pitch estimates of the noisy speech signal.

the CRLB of the DOA estimate in all situations, and the BH-MVDR method has a lower MSE than the WLS method in low SNRs (< 0 dB) and a small number of microphones ($M \leq 4$). In contrast, the proposed MVDR method has lower MSE than the competing methods, and reaches close to the CRLB in all situations.

In the joint DOA and pitch estimation, we apply the Capon 2D spectral density estimator [41] that is formulated for the spatio-temporal samples. We choose the length of sub-vectors of the temporal and spatial samples, respectively, as $\underline{N} = 20$ and $\underline{M} = 2$ in order to obtain a full-rank covariance matrix estimate of spatio-temporal samples. Fig. B.6 shows the MSE of the estimates in white and colored noise. As can be seen, the proposed MVDR method reaches the CRLB as well as the NLS method [26] in white noise, while the WLS method that is designed using the power spectrum estimates of the harmonics has a bias error. This error is because the Capon method suffers from the biased power spectrum estimate in small samples [42]. We can also see that the MVDR method reaches the bounds in different narrowband SNRs as opposed to the results of the NLS and the WLS methods in colored noise.

6.4 Real-Life Signal Analysis

In the first real-life experiment, we play back a female speech signal with a total length of 2.3 seconds along with a car noise signal in 10 dB SNR, and $f_s = 8.0$ kHz. The speech is an utterance of "Why were you away a year Roy?". We estimate the UFEs over 30 ms frames using the discrete Fourier

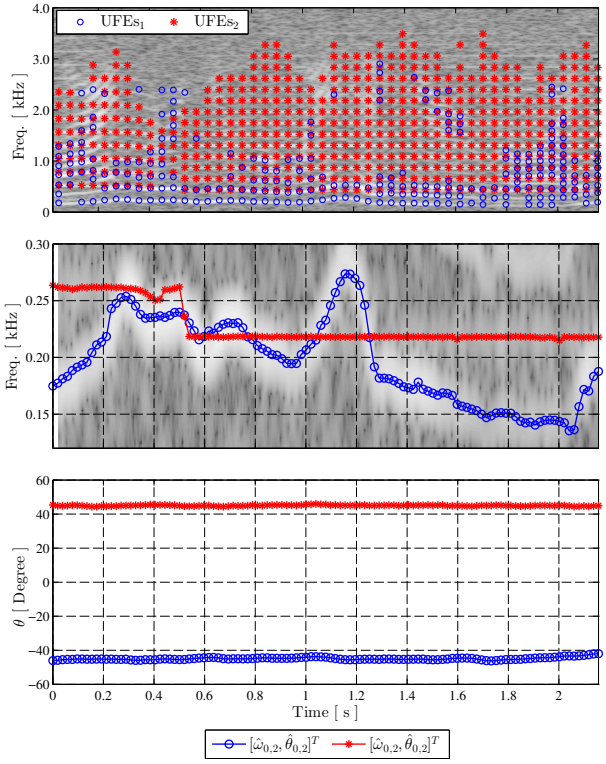


Fig. B.8: (top) Spectrogram and UFEs of the simultaneous female speaker and trumpet signals in 10 dB SNR of white noise. (middle) pitch and (bottom) DOA estimates of two sources at the same time.

transform (DFT) and the maximum a posteriori (MAP) model order estimator [36], and update the estimates at each time instance. We apply the time-varying mean value in (B.65) with $\lambda = 0.8$, and estimate the covariance matrix once per frame by averaging $B = 100$ earlier estimates. Fig. B.7 shows the spectrogram of the noisy signal as well as the UFEs, and the MVDR pitch estimates compared with the results of the WLS and the NLS pitch estimators. As one can see the NLS method has spurious estimates during 1.6 – 1.75 sec, and the WLS method has many spurious estimates because of the harmonics mismatch problem. In contrast, the proposed MVDR filtering method has a few outliers compared to the competing methods.

In another experiment we simulate an anechoic room, using the room impulse generator in [43], and play back the female speech signal along with a stepping down trumpet signal from the fixed directions $\theta_{0,1} = -\pi/4$ and $\theta_{0,2} = \pi/4$ radians, respectively, in 10 dB SNR of spatial white noise on $M = 5$ microphones of the ULA with $\delta = 0.04$ m. The output power of

Table B.1: MSE of joint pitch and DOA estimates in a room with reverberation

		Synthetic	Trumpet	Violin
$\hat{\omega}_0$	MVDR	5.1×10^{-9}	3.1×10^{-8}	4.3×10^{-8}
	WLS	7.7×10^{-9}	2.6×10^{-7}	7.9×10^{-8}
	NLS	5.7×10^{-10}	1.2×10^{-7}	9.4×10^{-8}
$\hat{\theta}_0$	MVDR	1.0×10^{-3}	3.0×10^{-3}	5.2×10^{-3}
	WLS	1.4×10^{-3}	1.5×10^{-3}	7.2×10^{-3}
	NLS	1.8×10^{-3}	5.5×10^{-3}	1.1×10^{-3}

the MVDR broadband beamformer [44] is used to estimate the 2D spectral density. The frequency and the corresponding DOA estimates are separated with respect to the initial DOA estimates $\pm 0.056\pi$ radians of two sources (see Alg. 1). Fig. B.8 shows the spectrogram of the noisy signal at the first microphone and the UFEs of two sources. This figure also shows joint pitch and DOA estimates using the proposed MVDR filtering method. The outliers are smeared out using the dynamic programming given in [45]. As one can see, the pitch of the trumpet signal is estimated successfully, however, its first harmonic is drowned in the noise and the first harmonic of the speech signal. Moreover, even though the fundamental frequencies of two signal sources change close of each other, or overlap in sometimes, their DOA and pitch are estimated successfully.

Finally, we conduct experiments using a multichannel audio database (SMARD) [46] with the configuration number 2010 of the database with five microphones (channels $17, 18, \cdots, 21$) of a ULA. Monophonic notes of synthetic, trumpet, and violin sounds have been played back through a loudspeaker and recorded by the ULA with 0 dB SNR of white noise in a room with a reverberation time of approximately 0.15 seconds. We compare the results of the joint DOA and pitch estimators with the measured true DOA and the estimated pitch of the clean signals using the ORTH method [6], and show their MSE in Table B.1. We have applied Alg. 2, and used the estimated $\mathbf{d}_L$ in the MVDR and the WLS methods. The table indicates that the estimates using the NLS, MVDR, and WLS methods are slightly close with low errors, which verifies that the methods are applicable to real-life signals.

6.5 Complexity

The computational complexity of the WLS pitch estimator has been investigated in [7] and compared with the NLS method. Floating-point operations per second (flops) of a particular computing has shown that the WLS is computationally much simpler than the NLS method [7]. Here we investigate

Table B.2: Computation complexity of the estimation methods.

ω_0	NLS	$\mathcal{O}\left\{\left[N(2L^2+1)+L^3+N^2(L+1)\right]N_{\omega_0}\right\}$
	MVDR	$P+\mathcal{O}\left[L^3+L^2(B+1)+L(2B+3)+3\right]$
	WLS	$P+\mathcal{O}\left(3L+2\right)$
		where $P=\mathcal{O}\left[f(N,1,N_\omega,1)\right]$
θ_0	NLS	$\mathcal{O}\left\{\left[M(2L^2+1)+L^3+N^2M^2(L+1)\right]N_\theta\right\}$
	MVDR	$\mathcal{O}\{M(L^3+2L^2N+LN)+L[M^3+(B+4)M^2+(2B+8)M+9]+BL^2+(2B+5)L+2\}$
	WLS	$\mathcal{O}\left[M(L^3+2L^2N+LN)+10LM+11L+1\right]$
$\begin{bmatrix}\omega_0\\ \theta_0\end{bmatrix}$	NLS	$\mathcal{O}\left\{\left[NM(2L^2+1)+L^3+N^2M^2(L+1)\right]N_{\omega_0}N_\theta\right\}$
	MVDR	$P+\mathcal{O}\left[8L^3+4L^2(B+1)+2L(2B+3)+2\right]$
	WLS	$P+\mathcal{O}\left(5L+2\right)$
		where $P=\mathcal{O}\left[f(N,M,N_\omega,N_\theta)\right]$

the computational complexity of the NLS method [26], and compare with the proposed methods using big $\mathcal{O}$ notation. Table B.2 shows the computational complexity of the methods, where N_{ω_0}, N_θ, and N_ω are the number of grids for the ranges of pitch, DOA, and the entire frequency spectrum, respectively, and $f(\cdot)$ is the complexity of a spectral density estimator. For example, the complexity of the DFT and the MUSIC methods are approximately $(2NM+1)N_\omega N_\theta$ and $\underline{N}^3\underline{M}^3+(\underline{N}^2\underline{M}^2+\underline{N}\underline{M})N_\omega N_\theta$, respectively, where $\underline{N}$ and $\underline{M}$ are the number of sub-vectors of temporal and spatial samples to estimate the correlation matrix of the spatio-temporal samples [25]. We examine the computational complexity of the compared methods as a function of the number of samples using $L=5$, $M=4$, $N=60$, $B=100$, $N_{\omega_0}=400$, $N_\omega=4000$, and $N_\theta=180$. In this example, the length of sub-vectors are the rounded half of the number of samples. We can see in Fig. B.9 that the WLS and the MVDR methods are computationally much simpler than the NLS method in different numbers of samples. Moreover, the most computation complexity of the WLS and the MVDR methods is the computation of the spectral density estimation, $\mathcal{O}\left[f(\cdot)\right]$.

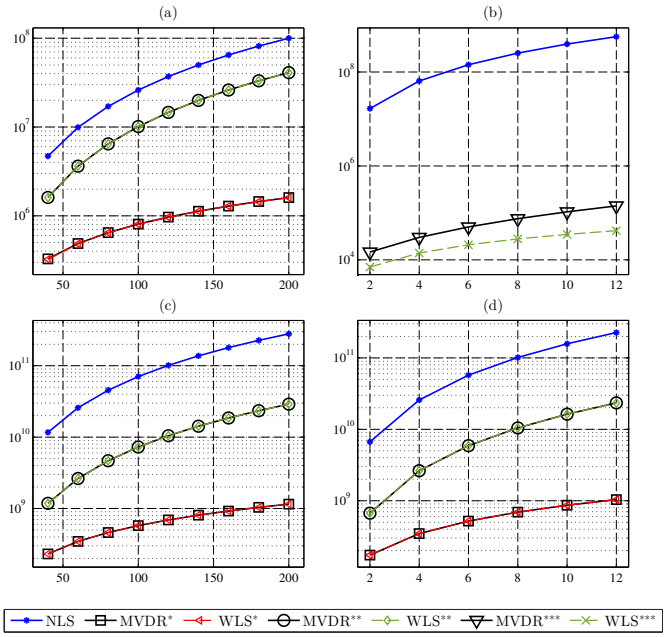


Fig. B.9: Computational complexity of (a) pitch estimation, (b) DOA estimation, and (c-d) joint pitch and DOA estimation. The superscripts * and ** denote the DFT and the MUSIC spectral density estimation methods, respectively, and *** denotes the two-steps DOA estimation methods.

7 Conclusion

In this paper, we have presented harmonic model-based DOA and pitch estimation methods associated with the distortionless constraints on the harmonic signal model. The proposed methods are designed using the narrowband noise statistics, which consequently makes the estimates robust against different types of noise. Without a detailed a priori assumption of the noise, we have estimated the narrowband noise statistics, at the frequencies of the harmonics, from the statistics of the parameter estimates of the harmonics. We have shown that the state-of-the-art and statistically efficient WLS pitch and DOA estimators are special cases of the proposed estimators in white Gaussian noise. Using a maximum-likelihood parameter estimator, we have also shown that the variance of the pitch and the DOA estimates are equal to the corresponding Cramér-Rao lower bounds, which are the smallest variance of the unbiased estimates. Experimental results also demonstrated the statistical efficiency of the proposed methods, and revealed that the other methods that are designed based on the wrong white noise assumption are suboptimal in colored noise. The results on real-life signals also confirmed

the applicability of the methods either in low local SNR or a scenario with overlapping frequencies of two sources. As a result, the proposed methods are robust in colored noise, and computationally simpler than the NLS DOA and pitch estimators.

References

[1] J. Benesty, S. Makino, and J. Chen, Eds., *Speech Enhancement*, ser. Signals and Communication Technology. Springer, 2005.

[2] A. Nehorai and B. Porat, "Adaptive comb filtering for harmonic signal enhancement," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 5, pp. 1124–1138, Oct. 1986.

[3] M. G. Christensen and A. Jakobsson, "Optimal filter designs for separating and enhancing periodic signals," *IEEE Trans. Signal Process.*, vol. 58, no. 12, pp. 5969–5983, Dec. 2010.

[4] T. Nakatani, M. Miyoshi, and K. Kinoshita, "Single-microphone blind dereverberation," in *Speech Enhancement*. Springer, 2005, pp. 247–270.

[5] M. G. Christensen and A. Jakobsson, "Multi-pitch estimation," *Synthesis Lectures on Speech and Audio Process.*, vol. 5, no. 1, pp. 1–160, 2009.

[6] M. G. Christensen, A. Jakobsson, and S. H. Jensen, "Joint high-resolution fundamental frequency and order estimation," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 5, pp. 1635–1644, Jul. 2007.

[7] H. Li, P. Stoica, and J. Li, "Computationally efficient parameter estimation for harmonic sinusoidal signals," *Elsevier Signal Process.*, vol. 80(9), pp. 1937–1944, 2000.

[8] M. Brandstein and D. Ward, Eds., *Microphone Arrays - Signal Processing Techniques and Applications*. Springer-Verlag, 2001.

[9] J. Benesty, Y. Huang, and J. Chen, *Microphone Array Signal Processing*. Springer-Verlag, 2008, vol. 1.

[10] M. S. Brandstein and H. Silverman, "A practical methodology for speech source localization with microphone arrays," *Comput. Speech Language*, 1997.

[11] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[12] J. Chen, J. Benesty, and Y. Huang, "Robust time delay estimation exploiting redundancy among multiple microphones," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 549–557, 2003.

[13] J. R. Jensen, J. K. Nielsen, M. G. Christensen, and S. H. Jensen, "On frequency domain models for TDOA estimation," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Sept. 2015.

[14] Y. Chan, R. Hattin, and J. B. Plant, "The least squares estimation of time delay and its use in signal detection," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 26, no. 3, pp. 217–222, Jun 1978.

[15] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, "Real-time multiple sound source localization and counting using a circular microphone array," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 21, no. 10, pp. 2193–2206, 2013.

[16] J. R. Jensen, M. G. Christensen, and S. H. Jensen, "Statistically efficient methods for pitch and DOA estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 3900–3904.

[17] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays - Signal Processing Techniques and Applications*, M. S. Brandstein and D. B. Ward, Eds. Springer-Verlag, 2001, ch. 8, pp. 157–180.

[18] M. Wohlmayr and M. Képesi, "Joint position-pitch extraction from multichannel audio," in *Proc. Interspeech*, Aug. 2007, pp. 1629–1632.

[19] J. M. Kates, "Classification of background noises for hearing-aid applications," *Journal of the Acoustical Society of America*, vol. 97, no. 1, pp. 461–470, 1995.

[20] J. H. McClellan, "Multidimensional spectral estimation," *Proc. IEEE*, vol. 70, no. 9, pp. 1029–1039, 1982.

[21] J. X. Zhang, M. G. Christensen, S. H. Jensen, and M. Moonen, "Joint DOA and multi-pitch estimation based on subspace techniques," *EURASIP J. on Advances in Signal Processing*, vol. 2012, no. 1, pp. 1–11, Jan. 2012.

[22] Y. Wu, L. Amir, J. R. Jensen, and G. Liao, "Joint pitch and doa estimation using the esprit method," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 23, no. 1, pp. 32–45, 2015.

[23] S. Karimian-Azari, J. R. Jensen, and M. G. Christensen, "Fast joint DOA and pitch estimation using a broadband MVDR beamformer," in *Proc. European Signal Processing Conf.*, Sept. 2013, pp. 1–5.

[24] J. Dmochowski, J. Benesty, and S. Affes, "Linearly constrained minimum variance source localization and spectral estimation," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 16, no. 8, pp. 1490–1502, Nov. 2008.

[25] J. R. Jensen, M. G. Christensen, J. Benesty, and S. H. Jensen, "Joint spatio-temporal filtering methods for DOA and fundamental frequency estimation," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 23, no. 1, pp. 174–185, 2015.

[26] J. R. Jensen, M. G. Christensen, and S. H. Jensen, "Nonlinear least squares methods for joint DOA and pitch estimation," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 21, no. 5, pp. 923 –933, May 2013.

[27] S. Tretter, "Estimating the frequency of a noisy sinusoid by linear regression (corresp.)," *IEEE Trans. Inf. Theory*, vol. 31, no. 6, pp. 832–835, 1985.

[28] M. G. Christensen, "Metrics for vector quantization-based parametric speech enhancement and separation," *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 3062–3071, 2013.

[29] S. L. Marple, Jr., "Computing the discrete-time "analytic" signal via FFT," *IEEE Trans. Signal Process.*, vol. 47, no. 9, pp. 2600–2603, Sep. 1999.

[30] M. G. Christensen, "Accurate estimation of low fundamental frequencies from real-valued measurements," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 21, no. 10, pp. 2042–2056, 2013.

[31] D. Tse and P. Viswanath, *Fundamentals of wireless communication*. Cambridge university press, 2005.

[32] P. Stoica and R. Moses, *Spectral Analysis of Signals*. Pearson Education, Inc., 2005.

[33] D. Rife and R. Boorstyn, "Single tone parameter estimation from discrete-time observations," *IEEE Trans. Inf. Theory*, vol. 20, no. 5, pp. 591–598, Sep 1974.

[34] M. H. Hayes, *Statistical digital signal processing and modeling*. John Wiley & Sons, 2009.

[35] J. Li and P. Stoica, "An adaptive filtering approach to spectral estimation and SAR imaging," *IEEE Trans. Signal Process.*, vol. 44, no. 6, pp. 1469–1484, Jun. 1996.

[36] P. Djuric, "A model selection rule for sinusoids in white gaussian noise," *IEEE Trans. Signal Process.*, vol. 44, no. 7, pp. 1744–1751, Jul 1996.

[37] P. Stoica and Y. Selen, "Model-order selection: a review of information criterion rules," *IEEE Signal Process. Mag.*, vol. 21, no. 4, pp. 36–47, Jul. 2004.

[38] S. Karimian-Azari, J. R. Jensen, and M. G. Christensen, "Fundamental frequency and model order estimation using spatial filtering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* IEEE, May 2014, pp. 5964–5968.

[39] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, Mar. 1986.

[40] K. Itoh, "Analysis of the phase unwrapping algorithm," *Applied Optics*, vol. 21, no. 14, pp. 2470–2470, 1982.

[41] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.

[42] A. Jakobsson and P. Stoica, "Combining capon and apes for estimation of spectral lines," *Circuits, Systems and Signal Processing*, vol. 19, no. 2, pp. 159–169, March 2000.

[43] E. A. P. Habets, "Room impulse response generator," Technische Universiteit Eindhoven, Eindhoven, Netherlands, Tech. Rep., 2010, ver. 2.0.20100920.

[44] M. E. Lockwood and et al., "Performance of time- and frequency-domain binaural beamformers based on recorded signals from real rooms," *J. Acoust. Soc. Am.*, vol. 115, no. 1, pp. 379–391, Jan. 2004.

[45] H. Ney, "A dynamic programming algorithm for nonlinear smoothing," *Signal Processing*, vol. 5, no. 2, pp. 163–173, March 1983.

[46] J. K. Nielsen, J. R. Jensen, S. H. Jensen, and M. G. Christensen, "The single-and multichannel audio recordings database (SMARD)," Sept. 2014, pp. 40–44.

Paper C

Multi-Pitch Estimation and Tracking using Bayesian Inference in Block Sparsity

Sam Karimian-Azari, Andreas Jakobsson, Jesper Rindom Jensen, and Mads Græsbøll Christensen

The paper has been published in the
Proceeding European Signal Processing Conf., pp. 16-20, 2015.

The layout has been revised.

Abstract

In this paper, we consider the problem of multi-pitch estimation and tracking of an unknown number of harmonic audio sources. The regularized least-squares is a solution for simultaneous sparse source selection and parameter estimation. Exploiting block sparsity, the method allows for reliable tracking of the found sources, without posing detailed a priori assumptions of the number of harmonics for each source. The method incorporates a Bayesian prior and assigns data-dependent regularization coefficients to efficiently incorporate both earlier and future data blocks in the tracking of estimates. In comparison with fix regularization coefficients, the simulation results, using both real and synthetic audio signals, confirm the performance of the proposed method.

1 Introduction

Estimation of the fundamental frequency, or pitch, detailing a set of audio sources, is an important problem in a wide range of applications, such as source separation, music transcription, and enhancement [1–3]. In speech recognition, for example, reliable pitch estimates are required in a prosodic implementation. The topic has for this reason attracted much interest, in particular for single pitch estimation [4], but the more challenging problem of multi-pitch estimation has also been given notable attention [5–8]. Often, these methods make strong *a priori* assumptions on the number of measured sources, as well as on the model orders of these sources. To determine such model order information is well known to be challenging [6], although some efforts on joint pitch and model order estimator techniques have been presented for the single pitch case [9]. For joint multi-pitch and model order estimation of the given number of sources, the problem have been formulated for polyphonic music transcription [5].

The recent pitch estimation using block sparsity (PEBS) technique introduced in [8] avoids such assumptions by imposing a verity of sparsity constraints, such that from a large dictionary of feasible pitches, both the number of sources and the model order of each found source can be determined. In this work, we introduce an extension of the PEBS algorithm to allow the efficient tracking of audio sources. Given the natural behavior of audio signals, the pitch often changes smoothly over time. That makes pitch values in sequential data frames highly correlated, which is often exploited in pitch tracking [10–12]. To allow for such temporal smoothness, we introduce data-dependent regularization coefficients for the sparsity constraints in the PEBS method, such that the estimate for the currently processed data frame is affected by the local spectral neighborhood of both the past and future data frames. The approach builds on earlier work on the adaptive Lasso [13] and

the Bayesian Lasso [14], as well as use a Gaussian smoothing kernel to regularize the corresponding components in the PEBS dictionary.

The remainder of this paper is organized as follows: In the next section, we present the signal model. In Section 3, we present the proposed multi-pitch estimation and tracking using Bayesian inference. Experimental results are presented in Section 4. Finally, we conclude on our work in Section 5.

2 Signal Model

Consider a sum of M harmonic audio sources, each with a fundamental frequency ω_m, and containing L_m harmonics, for $m = 1, 2, \ldots, M$, and Let

$$\mathbf{y}_n = \begin{bmatrix} y(n) & y(n+1) & \ldots & y(n+N-1) \end{bmatrix}^T \tag{C.1}$$

denote the data frame processed at time n, with N being the length of the frame. To simplify the notation and to reduce the resulting computational complexity, we here model the discrete-time analytical signal of the measured signals, as obtained using the method detailed in [15] (see also [6]). Thus, $\mathbf{y}_n$ may be well modeled as

$$\mathbf{y}_n \triangleq \sum_{m=1}^{M} \mathbf{Z}_m \mathbf{b}_m + \mathbf{v} = \mathbf{Z}\mathbf{b} + \mathbf{v} \tag{C.2}$$

where

$$\begin{aligned}
\mathbf{Z} &= \begin{bmatrix} \mathbf{Z}_1 & \mathbf{Z}_2 & \ldots & \mathbf{Z}_M \end{bmatrix} \\
\mathbf{Z}_m &= \begin{bmatrix} \mathbf{z}_m & \mathbf{z}_m^2 & \ldots & \mathbf{z}_m^{L_m} \end{bmatrix} \\
\mathbf{z}_m^l &= \begin{bmatrix} 1 & e^{jl\omega_m} & \ldots & e^{jl\omega_m(N-1)} \end{bmatrix}^T \\
\mathbf{b} &= \begin{bmatrix} \mathbf{b}_1^T & \mathbf{b}_2^T & \ldots & \mathbf{b}_M^T \end{bmatrix}^T \\
\mathbf{b}_m &= \begin{bmatrix} b_{m,1} & b_{m,2} & \ldots & b_{m,L_m} \end{bmatrix}^T
\end{aligned}$$

and $(\cdot)^T$ denotes the transpose. The matrix $\mathbf{Z}$ contains the $L_{\text{tot}} = \sum_{m=1}^{M} L_m$ complex-valued sinusoids, with the corresponding complex amplitudes $\mathbf{b}$, and is formed out of sub-basis matrices, $\mathbf{Z}_m$, detailing the tones presented in each of the M sources. The additive noise, $\mathbf{v}$, is here formed similar to $\mathbf{y}_n$ in (C.1), and is assumed to be a circularly symmetric Gaussian distributed white noise, i.e., $E\{\mathbf{v}(n)\mathbf{v}^H(n)\} = \sigma_v^2 \mathbf{I}_N$, where $E\{\cdot\}$ denotes the expectation.

3 Multi-pitch Estimation and Tracking

Consider the problem of spectral amplitude estimation of multiple sinusoids from the observed signal $\mathbf{y}_n$. For the given (known) basis matrix $\mathbf{Z}$, with

$N \gg L_{\text{tot}}$, and where the complex basis vectors $\mathbf{z}_m^l$ are assumed to be independent, one may form an estimate of the unknown pitch frequencies using the ordinary least-squares (LS) method, minimizing the sum of squared residuals such that $\hat{\mathbf{b}} = (\mathbf{Z}^H\mathbf{Z})^{-1}\mathbf{Z}^H\mathbf{y}_n$. However, such a solution requires knowledge of both the number of sources and the number of harmonics for each source. To avoid these assumptions, we define a (large) dictionary matrix over the considered range of frequencies, $\omega_r \in [\omega_{\min}, \omega_{\max}]$, and harmonics, such that the allowed number of harmonics for the dictionary elements $r = 1, 2, \cdots, S$ are limited to $L_r = \lfloor \pi/\omega_r \rfloor$, where $\lfloor \cdot \rfloor$ denotes the truncation operation to the nearest lower integer. Accordingly,

$$\mathbf{y}_n \triangleq \mathbf{W}\mathbf{a} + \mathbf{v} \tag{C.3}$$

where the $N \times S$ dictionary matrix is formed as

$$\mathbf{W} = \begin{bmatrix} \mathbf{Z}_1 & \mathbf{Z}_2 & \dots & \mathbf{Z}_S \end{bmatrix} \tag{C.4}$$

where $S \gg M$. The spectral amplitudes of the $L_{\text{ext}} = \sum_{r=1}^{S} L_r$ sinusoids of the dictionary, i.e.,

$$\mathbf{a} = \begin{bmatrix} \mathbf{a}_1^T & \mathbf{a}_2^T & \dots & \mathbf{a}_S^T \end{bmatrix}^T \tag{C.5}$$

are exceedingly sparse, containing only L_{tot} non-zero values. Then, for the problem of multi-pitch estimation, we form an estimate of the pitch frequencies by maximizing the likelihood of the spectral amplitude estimates, $\hat{\mathbf{a}}$, of the corresponding frequencies, such that

$$\hat{\mathbf{\Omega}} = \arg\max_{\mathbf{\Omega}} P\big(\{\|\hat{\mathbf{a}}_r\|_2\}_{r=1}^{S} \big| \mathbf{\Omega}\big) \tag{C.6}$$

where $\mathbf{\Omega} = \begin{bmatrix} \omega_1 & \omega_2 & \dots & \omega_{\tilde{M}} \end{bmatrix}^T$, for a given $\tilde{M}$, which may differ from the true number of sources, M.

Under the assumption of circularly symmetric Gaussian noise, the spectral amplitude estimates may be formed using the maximum likelihood (ML) method, such that

$$\hat{\mathbf{a}}_{\text{ML}} = \arg\max_{\mathbf{a}} \log P(\mathbf{y}_n \big| \mathbf{a}, \sigma_v) \tag{C.7}$$

where

$$P(\mathbf{y}_n|\mathbf{a}, \sigma_v) = \frac{1}{(2\pi\sigma_v^2)^{N/2}} \exp\Big(-\frac{1}{2\sigma_v^2}\|\mathbf{y}_n - \mathbf{W}\mathbf{a}\|_2^2\Big)$$

is the likelihood function, with $\|\cdot\|_2$ denoting the ℓ_2-norm. Given that the additive noise is assumed to be white, the resulting ML estimate coincides

with the standard LS estimate, and may thus be efficiently formed accordingly. However, in order to avoid over-fitting, one often instead forms the regularized LS estimate (see, e.g., [16]). The least absolute shrinkage and selection operator (Lasso) [17] is a well known regularized LS estimator that shrinks the sum of absolute values of the amplitudes toward zero. Imposing a Laplace distribution on the amplitudes, the likelihood for those may be expressed as [14]

$$P(a_{r,l_r}|\tau_{r,l_r},\sigma_v) = \frac{\tau_{r,l_r}}{2\sigma_v}\exp\Big(-\frac{\tau_{r,l_r}}{\sigma_v}|a_{r,l_r}|\Big). \tag{C.8}$$

Interpreting the Lasso as a Bayesian posteriori estimator, we express the probability of the spectral amplitudes, given the observations, and using the parameter vector $\mathbf{\Psi} = \big\{\bigcup_{r=1}^{S}\bigcup_{l_r=1}^{L_r}\tau_{r,l_r}\big\}$, as

$$\begin{aligned} P(\mathbf{a}|\mathbf{y}_n,\mathbf{\Psi},\sigma_v) &\propto \prod_{r=1}^{S}\prod_{l_r=1}^{L_r} P(\mathbf{y}_n|a_{r,l_r},\tau_{r,l_r},\sigma_v)P(a_{r,l_r}|\tau_{r,l_r},\sigma_v) \\ &\propto \exp\Big(-\frac{1}{2\sigma_v^2}\|\mathbf{y}_n-\mathbf{W}\mathbf{a}\|_2^2\Big)\prod_{r=1}^{S}\prod_{l_r=1}^{L_r}\exp\Big(-\frac{\tau_{r,l_r}}{\sigma_v}|a_{r,l_r}|\Big). \end{aligned} \tag{C.9}$$

As noted in [8], one may further include the group sparsity constraint to restrict the number of variable solutions (see also [18, 19]). Therefore, we extend on this notation by expressing the probability of the grouped variables, using the parameter vector $\mathbf{\Psi}_r = \big\{\bigcup_{l_r=1}^{L_r}\tau_{r,l_r}\big\}$, as

$$P(\mathbf{a}|\mathbf{y}_n,\mathbf{\Psi},\sigma_v) \propto \exp\Big(-\frac{1}{2\sigma_v^2}\|\mathbf{y}_n-\mathbf{W}\mathbf{a}\|_2^2\Big)\prod_{r=1}^{S}P(\mathbf{a}_r|\mathbf{\Psi}_r,\sigma_v) \tag{C.10}$$

where $P(\mathbf{a}_r|\mathbf{\Psi}_r,\sigma_v) \propto \exp\Big(-\frac{\|\mathbf{\Psi}_r\|_2}{\sigma_v}\|\mathbf{a}_r\|_2\Big)$. Herein, we take into consideration the spectral neighborhood as it evolves over time, such that

$$\begin{aligned} \hat{\mathbf{a}} &= \arg\max_{\mathbf{a}}\log P(\mathbf{a}|\mathbf{y}_n,\mathbf{\Psi},\sigma_v) \\ &= \arg\min_{\mathbf{a}}\frac{1}{2}\|\mathbf{y}_n-\mathbf{W}\mathbf{a}\|_2^2 + J \end{aligned} \tag{C.11}$$

where J denotes the imposed constraints, formed as

$$J = \|\boldsymbol{\psi}_{\mathrm{L}}\odot\mathbf{a}\|_1 + \sum_{r=1}^{S}\|\boldsymbol{\psi}_{\mathrm{GL},r}\|_2\|\mathbf{a}_r\|_2 \tag{C.12}$$

and with $\odot$ denoting the element-wise matrix product. To allow for the required sparsity constraints [8], the penalty term J involves both the ℓ_1-norm

penalty for the ordinary Lasso and the ℓ_2-norm penalty for the group-Lasso. The real-valued and non-negative regularization coefficients

$$\boldsymbol{\psi}_{\mathrm{L}} = \begin{bmatrix} \boldsymbol{\psi}_{\mathrm{GL},1}^T & \boldsymbol{\psi}_{\mathrm{GL},2}^T & \cdots & \boldsymbol{\psi}_{\mathrm{GL},S}^T \end{bmatrix}^T \tag{C.13}$$

$$\boldsymbol{\psi}_{\mathrm{GL},r} = \begin{bmatrix} \psi_{\mathrm{GL},r,1} & \psi_{\mathrm{GL},r,2} & \cdots & \psi_{\mathrm{GL},r,L_r} \end{bmatrix}^T \tag{C.14}$$

are assigned to the individual and grouped sinusoids, respectively, to make a trade-off between the residual and penalties. In [8], the PEBS estimator was formulated using common regularization coefficients for the two norms, such that

$$J = \lambda_{\mathrm{L}} \|\mathbf{a}\|_1 + \sum_{r=1}^{S} \lambda_{\mathrm{GL},r} \|\mathbf{a}_r\|_2 \tag{C.15}$$

where $\lambda_{\mathrm{L}} = \tau\sigma_v$ and $\lambda_{\mathrm{GL},r} = \tau\sigma_v\sqrt{L_r}$ with the common shrinkage coefficient τ.

As shown in [8], the resulting minimization may suffer from spurious estimates for weak signals and/or onsets, occasionally resulting in an overestimation of the model order. To reduce the occurrence of such spurious estimates, and to allow for a smooth spectral evaluation over frames, we in the following expand on the penalties in (C.15) to instead allow for more flexible penalty terms. In order to do so, we introduce adaptive weighting of the penalty terms in PEBS, using the notation of an adaptive Lasso, as introduced in [13], such that

$$\|\boldsymbol{\psi}_{\mathrm{GL},r}\|_2 = \frac{\hat{\sigma}_v}{(\|\tilde{\mathbf{a}}_r\|_2)^k} \tag{C.16}$$

$$\psi_{\mathrm{GL},r,l_r} = \frac{\hat{\sigma}_v}{(|\tilde{a}_{r,l_r}|)^k} \tag{C.17}$$

where $k > 0$ is a user defined parameter, and with the noise variance being estimated as $\hat{\sigma}_v \cong \|\mathbf{y}_n - \mathbf{W}\tilde{\mathbf{a}}\|_2$, and $\tilde{\mathbf{a}} = E\{\mathbf{a}|\boldsymbol{\Psi}, \sigma_v\}$, with $E\{\cdot\}$ denoting the expectation. The resulting adaptive penalty thereby offers a more flexible trade-off between the mean-squared error (MSE) and the bias. The introduced penalty is reminiscent of the iterative re-weighting adaptive Lasso [13], wherein the bias is similarly reduced by applying less shrinkage to the important predictors.

As the frequency content of most audio signals are piecewise smooth [20, 21], it is reasonable to model the dominant components in each frame as being close to those in the earlier and the following frames. Thus, the neighboring frames can be expected have nearly the same expectation of the absolute values, i.e., $E\{|\mathbf{a}(n+t)||\boldsymbol{\Psi}, \sigma_v\} \simeq E\{|\mathbf{a}(n)||\boldsymbol{\Psi}, \sigma_v\}$. In practice, one may apply time averaging over $2T+1$ initial estimates of $\mathbf{a}(n)$ to find an esti-

mate of the expectation at the time instance n, such that

$$E\{\mathbf{a}(n)\big|\mathbf{\Psi},\sigma_v\} \approx \frac{1}{2T+1}\sum_{t=-T}^{T}\hat{\mathbf{a}}(n+t)\odot\mathbf{h}(t) \tag{C.18}$$

where $\mathbf{h}(t)$ is a phase shift vector depending on the specific frequencies of the dictionary with unit absolute values, and where $\hat{\mathbf{a}}(n)$ denotes the estimated amplitude vector at time n, as obtained from the initialization or the earlier processed frames. For fast varying spectral content, as well as for poor initial or earlier spectral estimates, we include a spectral smoothing, formed using kernel regression. Here, we make use of the Nadaraya-Watson method introduced in [22], which use a monotonic decay over spectral neighborhood of the considered centroid, such that

$$\bar{a}_{r,l_r} = \frac{\sum_{g=1}^{S}\sum_{l_g=1}^{L_g} K_{\mathbf{\Sigma}}(\mathbf{x}_g-\mathbf{x}_r)\ \tilde{a}_{g,l_g}}{\sum_{g=1}^{S}\sum_{l_g=1}^{L_g} K_{\mathbf{\Sigma}}(\mathbf{x}_g-\mathbf{x}_r)} \tag{C.19}$$

where the kernel function is defined as

$$K_{\mathbf{\Sigma}}(\mathbf{x}_g-\mathbf{x}_r) = \exp\Big(-\frac{1}{2}(\mathbf{x}_g-\mathbf{x}_r)^T\mathbf{\Sigma}^{-1}(\mathbf{x}_g-\mathbf{x}_r)\Big)$$

with $\mathbf{\Sigma}$ denoting the diagonal covariance matrix, giving more weight to the amplitudes $\tilde{a}_{g,l_g}$ at the data point $\mathbf{x}_g = [\omega_g, l_g\omega_g]^T$ that has a smaller Euclidean distance to $\mathbf{x}_r = [\omega_r, l_r\omega_r]^T$.

4 Experimental Results

To investigate the performance of the extended PEBS method, we conducted simulations using both synthetic and real audio signals. Since the PEBS method preferably outperforms most stat-of-the-art methods, such as Capon, ANLS, and ORTH [8], we here only compare the found results with the PEBS method. In these simulations, we solved the convex minimization in (C.11) using the Matlab CVX package [23].

In the first experiment, we estimate the spectral amplitudes of a single-source synthetic signal for varying number of samples and signal-to-noise ratio (SNR). The synthetic signal was generated using the signal model in (C.2). The fundamental frequency of these signals were uniformly drawn on $\omega_1 \in [160, 290] \times (2\pi/f_s)$, with a uniformly distributed number of harmonics $L_1 \in \mathcal{U}\{5, \lfloor \pi/\omega_1 \rfloor\}$, unit amplitudes, and sampling frequency $f_s = 8.0$ kHz. The used dictionary contained $S = 130$ candidate pitches. The expectation in (C.16) and (C.17) was approximated using (C.18) with $k = 0.5$. Fig. C.1 shows the resulting normalized MSE as obtained from 100 Monte-Carlo simulations.

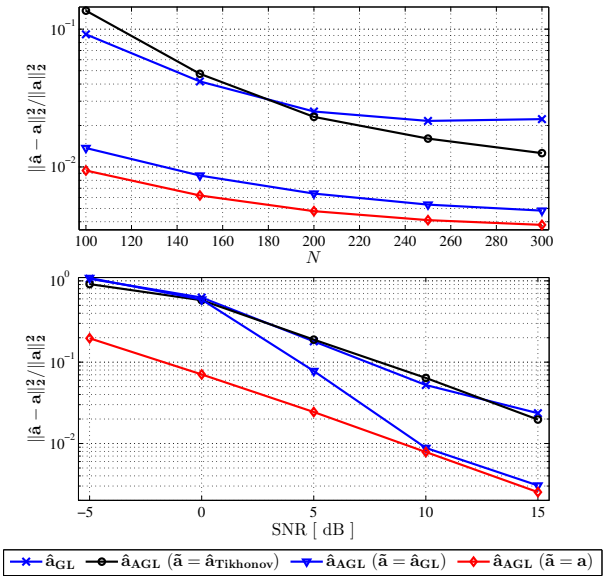


Fig. C.1: Normalized MSE of the spectral amplitude estimates versus the sample length N, at SNR $= 10$ dB (top), and versus SNR, using $N = 150$ (bottom).

As comparison, the figure shows the amplitude estimates of the PEBS method with the adaptive penalties, $\hat{\mathbf{a}}_{\mathrm{AGL}}$, using different initiation estimates: the PEBS amplitude estimates with common penalties ($\tilde{\mathbf{a}} = \hat{\mathbf{a}}_{\mathrm{GL}}$), the Tikhonov[1] amplitude estimates ($\tilde{\mathbf{a}} = \hat{\mathbf{a}}_{\mathrm{Tikh}}$), and the actual amplitudes ($\tilde{\mathbf{a}} = \mathbf{a}$). Here, the user parameters have been set as $\delta = 0.1$, $\lambda_{\mathrm{L}} = 0.12$, and $\lambda_{\mathrm{GL},r} = 0.12\sqrt{L_r}$. As is clear from the figure, the extended PEBS method offers an improved performance as compared to the regular PEBS algorithm, over all considered data lengths (except for the initial estimates using the Tikhonov estimator) and SNRs.

We proceed to examine a real audio signal consisting of a mixture of a female voice and a trumpet signal, corrupted by an additive white noise, with SNR $= 10$ dB, using $N = 150$ samples per frame. We apply $2T+1 = 3$ initial estimates in (C.18), using the regular PEBS estimates, and with $\boldsymbol{\Sigma} = \mathrm{diag}\{6.25, 0.01\} \times (2\pi/f_s)^2$ in the kernel smoother, where $\mathrm{diag}\{\cdot\}$ denotes a diagonal matrix formed from a vector argument. Fig. C.2 shows the spectrogram of the examined signal, together with the resulting pitch estimates of the two audio sources. As can be seen from the figure, the extended PEBS method estimates and tracks the audio sources smoothly, whereas the PEBS method suffer from some overshoots. For instance, at time 0.09 sec, the PEBS

[1]The Tikhonov estimator is formed as a regularized LS estimate such that $\hat{\mathbf{a}}_{\mathrm{Tikh}} = (\mathbf{W}^H\mathbf{W} + \delta\mathbf{I})^{-1}\mathbf{W}^H\mathbf{y}_n$, where $\delta \geq 0$ is the regularization coefficient, and $\mathbf{I} \in \mathbb{R}^{L_{\mathrm{ext}}}$ is an identity matrix.

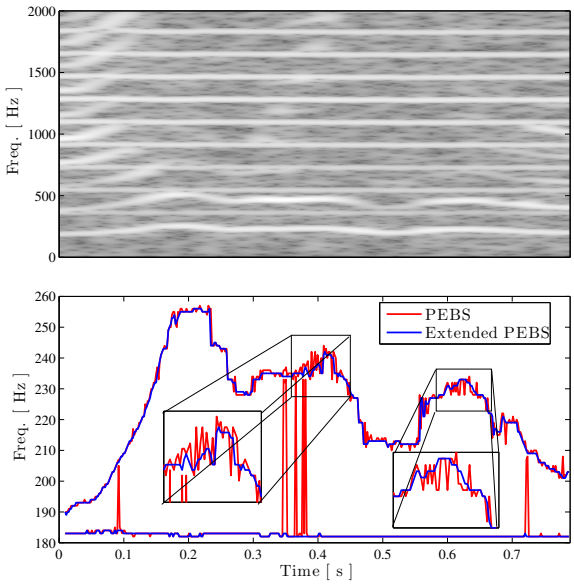


Fig. C.2: Spectrogram of the examined speech and trumpet signals (top), and the resulting multi-pitch estimates (bottom).

estimate finds the pitch of one of the sources close to the other, clearly mistakenly the spectral sidelobes of the first source for the pitch of the other signal source (see also Fig. 3). As can be seen from the Fig. C.3, the spectral amplitude estimates using the common PEBS method have some spurious non-zeros, and bias in comparison with the extended PEBS method.

5 Conclusion

In this work, we have presented a method for multi-pitch estimation and tracking of audio signals such as voiced speech and harmonic musical instruments, without assuming detailed prior knowledge about the signal sources. We have applied a general dictionary consisting of a set of groups for feasible fundamental frequencies and harmonics. Using ℓ_1-norm penalties is a well known solution for such the sparse signal formulation for both the individual and grouped sinusoids. We have shown that the regularization coefficients of the penalty terms should not be identical for all components of the dictionary, and assigned data-dependent regularization coefficients incorporated with an expectation on individual and grouped sinusoids. Experimental results have confirmed that the data-dependent regularization coefficients have

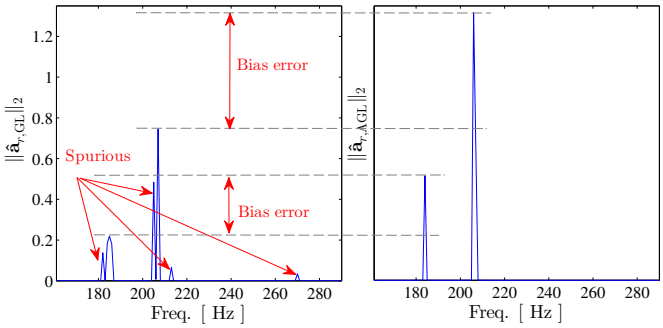


Fig. C.3: The ℓ_2-norm of spectral amplitude estimates using the common PEBS method (left), and the extended PEBS method (right).

a lower bias in comparison with the fixed ones. To track the pitch values smoothly over time, we have also applied a low-pass filter on the expected values to assign monotonic regularization coefficients regarding the spectral and temporal neighborhoods.

References

[1] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. IEEE Workshop Appl. of Signal Process. to Aud. and Acoust.* IEEE, 2003, pp. 177–180.

[2] B. Kostek, "Musical instrument classification and duet analysis employing music information retrieval techniques," *Proc. IEEE*, vol. 92, no. 4, pp. 712–729, 2004.

[3] J. Jensen and J. H. L. Hansen, "Speech enhancement using a constrained iterative sinusoidal model," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 7, pp. 731–740, Oct. 2001.

[4] L. Rabiner, M. Cheng, A. E. Rosenberg, and C. McGonegal, "A comparative performance study of several pitch detection algorithms," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 5, pp. 399–418, 1976.

[5] M. Davy, S. Godsill, and J. Idier, "Bayesian analysis of polyphonic western tonal music," *J. Acoust. Soc. Am.*, vol. 119, no. 4, pp. 2498–2517, Apr. 2006.

[6] M. G. Christensen and A. Jakobsson, "Multi-pitch estimation," *Synthesis Lectures on Speech and Audio Process.*, vol. 5, no. 1, pp. 1–160, 2009.

[7] H. Kameoka, T. Nishimoto, and S. Sagayama, "A multipitch analyzer based on harmonic temporal structured clustering," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 3, pp. 982–994, 2007.

[8] S. I. Adalbjörnsson, A. Jakobsson, and M. G. Christensen, "Multi-pitch estimation exploiting block sparsity," *Signal Processing*, vol. 109, pp. 236–247, 2015.

[9] M. G. Christensen, J. L. Højvang, A. Jakobsson, and S. H. Jensen, "Joint fundamental frequency and order estimation using optimal filtering," *EURASIP J. on Applied Signal Process.*, vol. 2011, no. 1, pp. 1–18, Jun. 2011.

[10] D. Talkin, "A robust algorithm for pitch tracking (rapt)," *Speech coding and synthesis*, vol. 495, p. 518, 1995.

[11] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 111, no. 4, pp. 1917–1930, Apr. 2002.

[12] H. Ney, "A dynamic programming algorithm for nonlinear smoothing," *Signal Processing*, vol. 5, no. 2, pp. 163–173, March 1983.

[13] H. Zou, "The adaptive lasso and its oracle properties," *Journal of the American statistical association*, vol. 101, no. 476, pp. 1418–1429, 2006.

[14] T. Park and G. Casella, "The bayesian lasso," *Journal of the American Statistical Association*, vol. 103, no. 482, pp. 681–686, 2008.

[15] S. L. Marple, Jr., "Computing the discrete-time "analytic" signal via FFT," *IEEE Trans. Signal Process.*, vol. 47, no. 9, pp. 2600–2603, Sep. 1999.

[16] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[17] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[18] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[19] J. Friedman, T. Hastie, and R. Tibshirani, "A note on the group lasso and a sparse group lasso," *arXiv preprint arXiv:1001.0736*, 2010.

[20] S. Karimian-Azari, N. Mohammadiha, J. R. Jensen, and M. G. Christensen, "Pitch estimation and tracking with harmonic emphasis on the acoustic spectrum," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, April 2015, pp. 4330–4334.

[21] C. Dubois and M. Davy, "Joint detection and tracking of time-varying harmonic components: a flexible bayesian approach," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 4, pp. 1283–1295, 2007.

[22] T. Hastie, R. Tibshirani, J. Friedman, T. Hastie, J. Friedman, and R. Tibshirani, *The elements of statistical learning*. Springer, 2009, vol. 2, no. 1.

[23] M. Grant and S. Boyd, "Matlab software for disciplined convex programming," *Online accessiable: http://cvxr.com/cvx*, 2013.

Paper D

Pitch Estimation and Tracking with Harmonic Emphasis on the Acoustic Spectrum

Sam Karimian-Azari, Nasser Mohammadiha, Jesper Rindom Jensen, and Mads Græsbøll Christensen

The paper has been published in the
Proceeding IEEE Int. Conf. Acoust., Speech, Signal Process. pp. 4330-4334, 2015.

Abstract

In this paper, we use unconstrained frequency estimates (UFEs) from a noisy harmonic signal and propose two methods to estimate and track the pitch over time. We assume that the UFEs are multivariate-normally-distributed random variables, and derive a maximum likelihood (ML) pitch estimator by maximizing the likelihood of the UFEs over short time-intervals. As the main contribution of this paper, we propose two state-space representations to model the pitch continuity, and, accordingly, we propose two Bayesian methods, namely a hidden Markov model and a Kalman filter. These methods are designed to optimally use the correlations in the consecutive pitch values, where the past pitch estimates are used to recursively update the prior distribution for the pitch variable. We perform experiments using synthetic data as well as a noisy speech recording, and show that the Bayesian methods provide more accurate estimates than the corresponding ML methods.

1 Introduction

Audio signals such as recordings of voiced speech and some music instruments can be modeled as a sum of harmonics with a fundamental frequency (or pitch). In practice, these signals are recorded in the presence of noise, and thus, the clean harmonic model will be less accurate. As a result, obtaining an accurate estimate of the pitch in noisy conditions is both challenging and very important for a wide range of applications such as enhancement, separation, and compression. Different pitch estimation methods have been investigated in [1, 2] which are based on a harmonic constraint. One common method to estimate the pitch is through the maximum likelihood (ML) framework [3]. In ML methods, consecutive pitch values are estimated independently, where obtaining a minimum-variance estimate is guaranteed [4, 5]. However, the pitch values in a sequence are usually highly correlated, which motivates the development of the Bayesian methods to optimally use the correlations. The Bayesian methods incorporate prior distributions, and can be used to derive the minimum mean square error (MMSE) estimator and the maximum a posteriori (MAP) estimator [6], e.g., [7].

State-of-the-art methods mostly track pitch estimates in a sequential process, e.g., [8–10]: first, pitch values are estimated in each time-frame, which is a sub-vector of the whole signal, and then they are smoothed, using a dynamic programming approach such as [11], without considering the noise statistics. For instance, the method in [8] uses a nonlinear smoothing method, which is a combination of median and low-pass filtering, and the method in [9] tracks pitch estimates based on a hidden Markov model (HMM). However, to obtain an optimal solution, the estimation and tracking have to be done jointly. One method that does this is proposed in [12], which operates

in the time-domain and uses a HMM based system to utilize the temporal correlation. This estimator is optimal if the noise is stationary with known statistics, while it is suboptimal in the more practical scenario where the noise statistics are unknown. A simple method to improve the performance in this scenario is to update the signal and noise statistics over time using a low-pass filter with exponential forgetting factor [13].

In this paper, we use the relation between harmonics to estimate and track the pitch in a harmonic signal. Herein, we jointly estimate and track pitch incorporating both the harmonic constraints and noise characteristics. First, we analytically find an optimal ML pitch estimator in each time-frame using unconstrained frequency estimates (UFEs)[1], which are the perturbed frequencies of harmonics in Gaussian noise [20]. One of the key contributions of this work is to transfer the pitch estimation problem with the harmonic constraints into a state-space representation where the state equation is designed to model the pitch evolution. Consequently, we can use a state-of-the-art Bayesian method to estimate the pitch values. We propose a discrete state-space representation, an HMM, using which we develop a MAP estimator for the pitch. We also propose a continuous state-space, a Kalman filter (KF), which is used to obtain an MMSE estimate of the pitch. Both the HMM and KF based methods utilize the correlations and lead to recursive pitch estimates.

The rest of this paper is organized as follows: In Section 2, we present the signal model, and introduce the ML pitch estimator. For a sequence of observations, the Bayesian estimators are presented in Section 3. Then, in Section 4, some experimental results are presented. In closing, the work is concluded in Section 5.

2 Problem Formulation

2.1 Signal Model

We model a harmonic signal[2], e.g., voiced speech, as a sum of $L(n)$ sinusoids at the time instance n like

$$s(n) = \sum_{l=1}^{L(n)} \alpha_l \, e^{j\,(\omega_l(n)n+\varphi_l)}, \tag{D.1}$$

[1] UFEs are multiple single-frequency tones, which are the location of peaks of spectral densities over frequency, assuming that the number of harmonics are known, e.g. using a method in [14, 15]. Different methods for estimation of the spectral density have been investigated in [16], e.g., using discrete Fourier transform (DFT), MUSIC [17], NLS [18], and Capon [19].

[2] Here, we utilize the discrete-time analytical signal, as in [21], to simplify the notation and reduce the resulting complexity.

where $\omega_l(n) = l\omega_0(n)$, and α_l and φ_l are amplitude and initial phase of each sinusoid, respectively. In the signal sub-vector $\mathbf{s}(n) = [\, s(n), s(n-1), \cdots, s(n-M-1)\,]^T$, we assume that the signal parameters are approximately stationary, and collect the constrained frequencies like

$$\mathbf{\Omega}(n) = [\, \omega_1(n),\, \omega_2(n),\, \ldots, \omega_L(n)\,]^T = \mathbf{d}_L(n)\,\omega_0(n), \tag{D.2}$$

where the superscript ^{T} is the transpose operator, and $\mathbf{d}_L(n) = [\, 1, 2, \ldots, L(n)\,]^T$. We assume that the harmonic signal $s(n)$ is contaminated by additive Gaussian noise $v(n)$ with the variance of σ^2 and zero mean as

$$x(n) = s(n) + v(n), \tag{D.3}$$

i.e., $v(n) \sim \mathcal{N}(0, \sigma^2)$. If the narrowband signal-to-noise ratios (SNRs) of sinusoids are high enough, the observed signal of such harmonic model can be approximated by the angular noise $\Delta\omega_l(n)$ with a zero-mean normal distribution on each sinusoid [22] as

$$x(n) \approx \sum_{l=1}^{L(n)} \alpha_l\, e^{j\,(\omega_l(n)n + \Delta\omega_l(n) + \varphi_l)}. \tag{D.4}$$

Therefore, unconstrained frequency estimates (UFEs)—of the constrained frequencies—can be approximated as the summation of the true frequencies and an error term $\Delta\mathbf{\Omega}(n)$ that is defined as $\Delta\mathbf{\Omega}(n) = [\, \Delta\omega_1(n), \Delta\omega_2(n), \ldots, \Delta\omega_L(n)\,]^T$ [20], i.e.,

$$\begin{aligned} \hat{\mathbf{\Omega}}(n) &= [\, \hat{\omega}_1(n),\, \hat{\omega}_2(n),\, \ldots,\, \hat{\omega}_L(n)\,]^T \\ &= \mathbf{\Omega}(n) + \Delta\mathbf{\Omega}(n), \end{aligned} \tag{D.5}$$

where $\Delta\mathbf{\Omega}(n)$ is a zero-mean multivariate-normally-distributed variable with the covariance matrix defined as

$$\mathbf{R}_{\Delta\mathbf{\Omega}}(n) = \mathrm{E}\{\Delta\mathbf{\Omega}(n)\Delta\mathbf{\Omega}^T(n)\}, \tag{D.6}$$

where $\Delta\mathbf{\Omega}(n) = \hat{\mathbf{\Omega}}(n) - \mathrm{E}\{\hat{\mathbf{\Omega}}(n)\}$, and $\mathrm{E}\{\cdot\}$ denotes the mathematical expectation. In white Gaussian noise, the precision matrix (inverse of the covariance matrix) is given by [20]:

$$\mathbf{R}^{-1}_{\Delta\mathbf{\Omega}}(n) = \frac{2}{\sigma^2}\,\mathrm{diag}\Big\{\, \alpha_1^2,\, \alpha_2^2,\, \ldots,\, \alpha_L^2 \,\Big\}, \tag{D.7}$$

where $\mathrm{diag}\{\cdot\}$ denotes the diagonal matrix formed with the vector input along its diagonal. Consequently, for the time frame $\mathbf{x}(n) = [\, x(n), x(n-1), \cdots, x(n-M-1)\,]^T$, the probability density function (PDF) of the UFEs given the unknown pitch is approximately given by a multivariate normal distribution with the constrained and non-zero mean:

$$P(\hat{\mathbf{\Omega}}(n) | \omega_0(n)) \sim \mathcal{N}(\mathbf{d}_L(n)\,\omega_0(n), \mathbf{R}_{\Delta\mathbf{\Omega}}(n)). \tag{D.8}$$

2.2 ML Pitch Estimate

Assuming that pitch is a deterministic parameter, the maximum likelihood (ML) estimator can be used to obtain an estimate for the pitch, where the log-likelihood function of the UFEs is maximized:

$$\hat{\omega}_0(n) = \arg\max_{\omega_0(n)} \log P(\hat{\mathbf{\Omega}}(n)|\omega_0(n)). \tag{D.9}$$

The optimal ML pitch estimator can be obtained by taking the first derivative of the likelihood function with respect to $\omega_0(n)$ and setting it to zero, and is given by

$$\hat{\omega}_0(n) = \left[\mathbf{d}_L^T(n)\mathbf{R}_{\Delta\mathbf{\Omega}}^{-1}(n)\mathbf{d}_L(n)\right]^{-1}\mathbf{d}_L^T(n)\mathbf{R}_{\Delta\mathbf{\Omega}}^{-1}(n)\hat{\mathbf{\Omega}}(n).$$

In the particular case with white Gaussian noise, the ML pitch estimator is simplified to

$$\hat{\omega}_0(n) = \frac{1}{\sum_{l=1}^{L}(l\ \alpha_l)^2}\ [\,\alpha_1^2,\, 2\alpha_2^2,\, \ldots,\, L\alpha_L^2\,]\,\hat{\mathbf{\Omega}}(n), \tag{D.10}$$

which is the same result as the weighted least squared (WLS) pitch estimator in [5].

3 Pitch Tracking

In general, the ML estimator is interesting because it is the minimum-variance unbiased estimator in Gaussian noise. Using M samples of a stationary signal, the minimum variance of the ML pitch estimator is inversely proportional to M^3 [1]. Speech signals generally are not stationary, but a voiced speech signal often has an stationary pitch during a short-time frame less than 30 ms that, consequently, limits the number of samples and the variance of the obtained pitch estimate. Moreover, pitch values are usually correlated in a sequence; this a priori information can be used to minimize the estimation error, which is the aim of this section.

In the following subsections, we compute the likelihood of a given $\hat{\mathbf{\Omega}}(n)$ using (D.8), for which we need to compute the covariance matrix using (D.6). To evaluate (D.6), the expected value $\mathrm{E}\{\hat{\mathbf{\Omega}}(n)\}$ has to be computed first. Since the pitch is varying over time, we use an exponential moving average (EMA) method with a forgetting factor $0 < \lambda < 1$ to recursively update the time-varying mean value as:

$$\mathrm{E}\{\hat{\mathbf{\Omega}}(n)\} = \lambda\,\hat{\mathbf{\Omega}}(n) + (1-\lambda)\,\mathrm{E}\{\hat{\mathbf{\Omega}}(n-1)\}. \tag{D.11}$$

After computing $\mathrm{E}\{\hat{\mathbf{\Omega}}(n)\}$, we can compute $\mathbf{R}_{\Delta\mathbf{\Omega}}(n)$ using (D.6). For this purpose, we use an ML estimator for the covariance (from normally-distributed observations) among N estimates [6]:

$$\mathbf{R}_{\Delta\mathbf{\Omega}}(n) = \frac{1}{N}\sum_{i=n-N+1}^{n} \Delta\mathbf{\Omega}(i)\Delta\mathbf{\Omega}^T(i). \tag{D.12}$$

3.1 Discrete State-Space: HMM

In this section, we assume that pitch is a discrete random variable and develop an HMM-based pitch estimation method to utilize the correlation between consecutive pitch values. For our problem, the hidden state corresponds to the pitch. HMM provides a simple and yet effective way to model the temporal correlations and has been widely used in speech processing [9, 23]. We discretize the interval that encloses the possible values of pitch into N_d centroids. In practice, since pitch is a continuous variable, the discretization may introduce a systematic bias in the estimation. However, this bias can be arbitrarily lowered by increasing N_d.

We use a first-order Markov model, where the state variable depends only on the one step past as:

$$P(\omega_0(n)|\omega_0(n-1),\cdots) = P(\omega_0(n)|\omega_0(n-1)), \tag{D.13}$$

where $P(\omega_0(n)|\omega_0(n-1))$ denotes the transition probability from $\omega_0(n-1)$ to $\omega_0(n)$, and $\sum_{\omega_0(n)} P(\omega_0(n)|\omega_0(n-1)) = 1$. By gathering all these probabilities, we obtain an $N_d \times N_d$ matrix which is usually referred to as the transition matrix. Since the neighboring pitch values are highly correlated, it is reasonable to assume that $\omega_0(n)$ is likely to be close to $\omega_0(n-1)$, and the probability of a pitch estimate far from $\omega_0(n-1)$ will be very small. In order to use this a priori information, we pre-define the transition matrix by sampling from a normal PDF. Hence, the diagonal elements of the transition matrix correspond to the maximum value of a normal PDF with the variance σ_t^2, and the neighboring values are sampled from the normal PDF in steps of one standard deviation.

In a hidden state-space model, we have a series of observations, i.e., UFEs, which indirectly relate to states, and each state has an emission distribution that is the same as the likelihood function in (D.8). We aim to estimate pitch (the hidden state) in a causal manner, i.e., given only the current and past observations $\{\hat{\mathbf{\Omega}}(n), \hat{\mathbf{\Omega}}(n-1), \ldots\}$. This yields a MAP estimate for pitch, and

the common method to implement it is through the forward algorithm [23]:

$$\hat{\omega}_0(n) = \arg\max_{\omega_0(n)} \log P(\omega_0(n)|\hat{\mathbf{\Omega}}(n), \hat{\mathbf{\Omega}}(n-1), \cdots) \tag{D.14}$$

$$= \arg\max_{\omega_0(n)} \log P(\hat{\mathbf{\Omega}}(n)|\omega_0(n)) + \log P(\omega_0(n)|\hat{\mathbf{\Omega}}(n-1), \hat{\mathbf{\Omega}}(n-2), \cdots), \tag{D.15}$$

that maximizes the log-likelihood function plus the logarithm of the prior distribution, which appears as a regularization term. The prior distribution is recursively updated as

$$P(\omega_0(n)|\hat{\mathbf{\Omega}}(n-1), \hat{\mathbf{\Omega}}(n-2), \cdots) = \sum_{\omega_0(n-1)} P(\omega_0(n)|\omega_0(n-1)) P(\omega_0(n-1)|\hat{\mathbf{\Omega}}(n-1), \cdots). \tag{D.16}$$

Note that the maximization in (D.14) is simply choosing the maximum value in an N_d-dimensional vector.

3.2 Continuous State-Space: Kalman Filter (KF)

As it was discussed in Section 3.1, pitch is a continuous variable and, hence, it is theoretically preferred to model the variations of pitch using a continuous state-space representation, e.g., [24]. In this section, we develop such model, where the state-evolution equation is designed to take into account the correlation of the pitch values in the consecutive frames. For this purpose, we write the complete state-space representation as follows:

$$\omega_0(n) = \omega_0(n-1) + \delta(n),$$
$$\hat{\mathbf{\Omega}}(n) = \mathbf{d}_L(n)\,\omega_0(n) + \Delta\mathbf{\Omega}(n),$$

where $\delta(n) \sim \mathcal{N}(0, \sigma_t^2)$ and $\Delta\mathbf{\Omega}(n) \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_{\Delta\mathbf{\Omega}}(n))$ are the state and observation noise, respectively, which are assumed to be independent. Kalman filtering is a well-known method that computes the MMSE estimate of the hidden state variable in above [25], which is used here.

First, a pitch estimate is predicted using the past estimates as

$$\hat{\omega}_0(n|n-1) = \hat{\omega}_0(n-1|n-1) \tag{D.17}$$

where $\hat{\omega}_0(n|n-1)$ denotes the predicted estimate using the past observations until $\hat{\mathbf{\Omega}}(n-1)$, and $\hat{\omega}_0(n-1|n-1)$ denotes the updated estimate at time $n-1$ using all the past observations, including $\hat{\mathbf{\Omega}}(n-1)$. The variance of the prediction is also given by

$$\sigma_\kappa^2(n|n-1) = \sigma_\kappa^2(n-1|n-1) + \sigma_t^2, \tag{D.18}$$

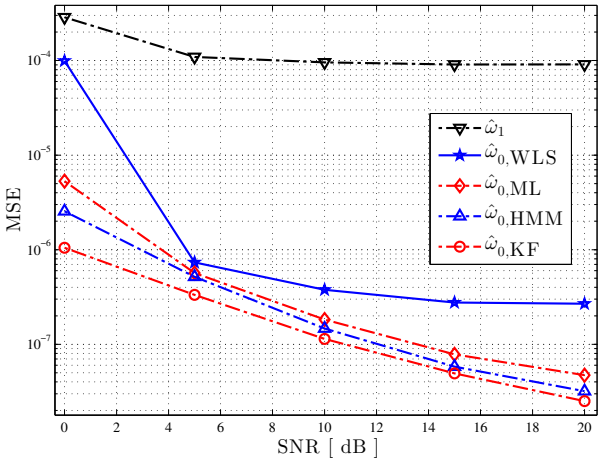


Fig. D.1: Obtained MSE using the proposed methods as a function of SNR. See text for details.

where $\sigma_{\kappa}^2(n|n-1)$ and $\sigma_{\kappa}^2(n-1|n-1)$ denote the variance of the predicted estimate and updated estimate, respectively.

Second, the pitch estimate is updated. For this purpose, the error term (or innovation) is computed as

$$\mathbf{e}(n) = \hat{\mathbf{\Omega}}(n) - \mathbf{d}_L(n)\,\hat{\omega}_0(n|n-1). \tag{D.19}$$

Then, the predicted estimate is updated:

$$\hat{\omega}_0(n|n) = \hat{\omega}_0(n|n-1) + \mathbf{h}_{\kappa}(n)\mathbf{e}(n), \tag{D.20}$$

where $\mathbf{h}_{\mathrm{K}}(n)$ denotes the Kalman gain and is given by

$$\mathbf{h}_{\kappa}(n) = \sigma_{\kappa}^2(n|n-1)\mathbf{d}_L^T(n)\left[\mathbf{\Pi}_L(n)\sigma_{\kappa}^2(n|n-1) + \mathbf{R}_{\Delta\mathbf{\Omega}}(n)\right]^{-1},$$

where $\mathbf{\Pi}_L(n) = \mathbf{d}_L(n)\mathbf{d}_L^T(n)$. The variance of the updated estimate is also recursively updated using

$$\sigma_{\kappa}^2(n|n) = \left[1 - \mathbf{h}_{\kappa}^T(n)\mathbf{d}_L(n)\right]\sigma_{\kappa}^2(n|n-1). \tag{D.21}$$

4 Experiment Results

We perform simulations to estimate and track the pitch in synthetic and real speech signals using the proposed methods. In the first experiment, we estimate the frequency of a sinusoid signal with the sampling frequency $f_s = 8.0$ kHz. A 65536-point discrete Fourier transform (DFT) was applied on data samples during 10 ms, i.e., $M = 80$. The forgetting factor λ in (D.11) was

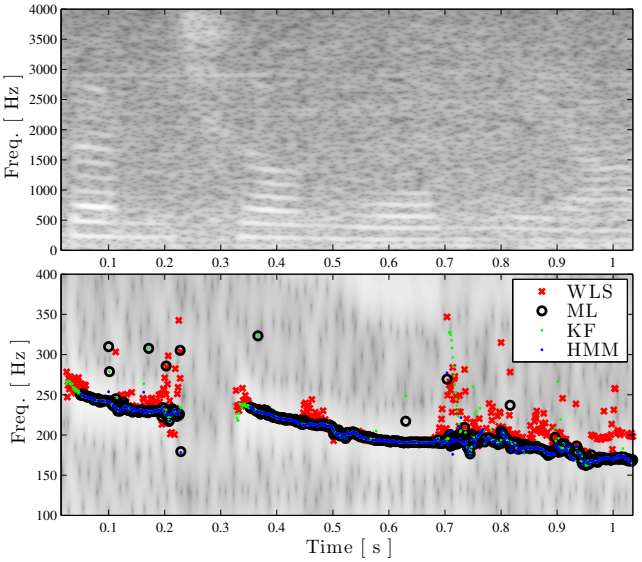


Fig. D.2: Spectrogram of a speech signal in the presence of car noise at SNR = 5 dB (top), and estimated pitch values, superimposed on the spectrogram (bottom).

set to 0.6, and $N = 50$ observations were used to estimate the noise covariance matrix in (D.12). The sinusoid signal in this experiment was a linear chirp signal with $L = 5$ harmonics with random phases and identical amplitudes during 0.1 s, which was then perturbed by additive white Gaussian noise at various signal-to-noise ratios (SNRs). The starting pitch of the chirp signal was $400\pi/f_s$ and it increases with a rate of $r = 100$ Hz/s. For the HMM-based pitch estimator, the frequency range $\omega \in [150, 280] \times (2\pi/f_s)$ was discretized into $N_d = 1000$ samples. The variance related to the state transition for both HMM- and KF-based methods was set to be proportional to the linear chirp rate, i.e., $\sigma_t = \sqrt{2}\pi r/f_s^2$. Fig. D.1 shows the obtained Mean Square Error (MSE), using 100 Monte-Carlo simulations for each SNR. As can be seen, the HMM- and KF-based pitch estimates have lower MSE than the corresponding ML pitch estimate, $\hat{\omega}_{0,\mathrm{ML}}$, and a state-of-the-art pitch estimator from [5], which is denoted by $\hat{\omega}_{0,\mathrm{WLS}}$. Moreover, the figure shows that the first harmonic of the UFEs (denoted by $\hat{\omega}_1$) results in significantly larger errors than all the other methods.

In the next experiment, we estimate the pitch in a speech signal degraded by car noise at SNR = 5 dB. We select voiced speech segments using the normalized low frequency energy ratio [26], and estimate the number of harmonics using the MAP order estimation [15]. A fixed $\sigma_t = 0.0318\pi/f_s$ was used for both HMM- and KF-based methods. The other parameters were set: $M = 240$, $\lambda = 0.9$, and $N = 150$, as the best choise for this experiment.

Fig. G.3 depicts the estimated pitch values on the spectrogram of the noisy signal. As can be observed, the HMM-based method tracks the pitch values smoothly and more accurately compared to the other methods.

5 Conclusion

The work presented in this paper has focused on pitch estimation. We have formulated an ML estimator for the pitch, which was then extended to utilize the correlations between consecutive pitch values to achieve higher accuracy and continuity for sequential pitch estimates. We have proposed HMM- and KF-based pitch estimation methods from the unconstrained frequency estimates, where noise characteristics were updated recursively. These characteristics make a contour over the frequency and time evolution, which were considered in the joint pitch estimation and tracking. Experimental results showed that both HMM- and KF-based methods outperform the corresponding optimal ML pitch estimator and another state-of-the-art method, based on the weighted least squares. Moreover, results using a real speech signal showed that the HMM-based method tracks the pitch more accurately and smoothly than the KF-based method.

References

[1] M. G. Christensen and A. Jakobsson, "Multi-pitch estimation," *Synthesis Lectures on Speech and Audio Process.*, vol. 5, no. 1, pp. 1–160, 2009.

[2] W. J. Hess, "Pitch and voicing determination of speech with an extension toward music signals," *Springer Handbook of Speech Processing*, pp. 181–212, 2008.

[3] D. Rife and R. Boorstyn, "Single tone parameter estimation from discrete-time observations," *IEEE Trans. Inf. Theory*, vol. 20, no. 5, pp. 591–598, Sep 1974.

[4] P. Stoica, A. Jakobsson, and J. Li, "Cisoid parameter estimation in the colored noise case: asymptotic Cramér-Rao bound, maximum likelihood, and nonlinear least-squares," *IEEE Trans. Signal Process.*, vol. 45, no. 8, pp. 2048–2059, Aug. 1997.

[5] H. Li, P. Stoica, and J. Li, "Computationally efficient parameter estimation for harmonic sinusoidal signals," *Elsevier Signal Process.*, vol. 80(9), pp. 1937–1944, 2000.

[6] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice Hall, Inc., 1993.

[7] L. Parra and U. Jain, "Approximate kalman filtering for the harmonic plus noise model," in *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the*. IEEE, 2001, pp. 75–78.

[8] L. Rabiner, M. Sambur, and C. Schmidt, "Applications of a nonlinear smoothing algorithm to speech processing," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 23, no. 6, pp. 552–557, 1975.

[9] M. Wohlmayr, M. Stark, and F. Pernkopf, "A probabilistic interaction model for multipitch tracking with factorial hidden markov models," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 19, no. 4, pp. 799–810, 2011.

[10] D. Talkin, "A robust algorithm for pitch tracking (rapt)," *Speech coding and synthesis*, vol. 495, p. 518, 1995.

[11] H. Ney, "A dynamic programming algorithm for nonlinear smoothing," *Signal Processing*, vol. 5, no. 2, pp. 163–173, March 1983.

[12] J. Tabrikian, S. Dubnov, and Y. Dickalov, "Maximum a-posteriori probability pitch tracking in noisy environments using harmonic model," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 1, pp. 76 – 87, Jan. 2004.

[13] M. G. Christensen, "A method for low-delay pitch tracking and smoothing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2012, pp. 345–348.

[14] P. Stoica and Y. Selen, "Model-order selection: a review of information criterion rules," *IEEE Signal Process. Mag.*, vol. 21, no. 4, pp. 36–47, Jul. 2004.

[15] P. M. Djuric, "Asymptotic MAP criteria for model selection," *IEEE Trans. Signal Process.*, vol. 46, no. 10, pp. 2726 –2735, Oct. 1998.

[16] P. Stoica and R. Moses, *Spectral Analysis of Signals*. Pearson Education, Inc., 2005.

[17] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, Mar. 1986.

[18] P. Stoica and A. Nehorai, "Statistical analysis of two non-linear least-squares estimators of sine waves parameters in the colored noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1988, pp. 2408–2411 vol.4.

[19] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.

[20] S. Karimian-Azari, J. R. Jensen, and M. G. Christensen, "Robust pitch estimation using an optimal filter on frequency estimates," in *Proc. European Signal Processing Conf.*, Sept . 2014, pp. 1557–1561.

[21] S. L. Marple, Jr., "Computing the discrete-time "analytic" signal via FFT," *IEEE Trans. Signal Process.*, vol. 47, no. 9, pp. 2600–2603, Sep. 1999.

[22] S. Tretter, "Estimating the frequency of a noisy sinusoid by linear regression (corresp.)," *IEEE Trans. Inf. Theory*, vol. 31, no. 6, pp. 832–835, 1985.

[23] L. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[24] C. Dubois and M. Davy, "Joint detection and tracking of time-varying harmonic components: a flexible bayesian approach," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 4, pp. 1283–1295, 2007.

[25] T. Kailath, A. H. Sayed, and B. Hassibi, *Linear Estimation*. Prentice Hall, 2000.

[26] S. A. Zahorian and H. Hu, "A spectral/temporal method for robust fundamental frequency tracking," *J. Acoust. Soc. Am.*, vol. 123, no. 6, pp. 4559–4571, 2008.

Paper E

Fast Joint DOA and Pitch Estimation using a Broadband MVDR Beamformer

Sam Karimian-Azari, Jesper Rindom Jensen,
and Mads Græsbøll Christensen

The paper has been published in the
Proceeding European Signal Processing Conf., pp. 1–5, 2013.

The layout has been revised.

Abstract

The harmonic model, i.e., a sum of sinusoids having frequencies that are integer multiples of the pitch, has been widely used for modeling of voiced speech. In microphone arrays, the direction-of-arrival (DOA) adds an additional parameter that can help in obtaining a robust procedure for tracking non-stationary speech signals in noisy conditions. In this paper, a joint DOA and pitch estimation (JDPE) method is proposed. The method is based on the minimum variance distortionless response (MVDR) beamformer in the frequency-domain and is much faster than previous joint methods, as it only requires the computation of the optimal filters once per segment. To exploit that both pitch and DOA evolve piece-wise smoothly over time, we also extend a dynamic programming approach to joint smoothing of both parameters. Simulations show the proposed method is much more robust than parallel and cascaded methods combining existing DOA and pitch estimators.

1 Introduction

The estimation of the fundamental frequency, or pitch as it is commonly referred to, of voiced speech signals is a challenging problem, as it is a key feature in many solutions for enhancement, separation, classification, compression, coding, etc. Various methods have been investigated to solve this problem for the single-channel case (see, e.g., [1–3]). However, for the multi-channel case, much less work has been done. When using a microphone array, the spatial information, such as the direction-of-arrival (DOA), provides additional information that can be used for such tasks as separation and enhancement. Likewise, a spatial filter, or beamformer, can be used for extracting signals impinging on the array from any particular DOA [4]. Multiple concurrent speech signals, which each are of broadband nature, may have overlapping spectral features with common pitch and harmonics and are, hence, difficult to separate. A beamforming technique can be used for locating and separating such signals by joint estimation of both the DOA and pitch of the desired signal in cases where it would otherwise not be possible.

The estimation of each parameter has traditionally been treated separately [5], and estimates of both can be obtained using cascaded or parallel combinations of pitch and DOA estimators. In the cascaded approach, a broadband beamformer estimates the DOA and extracts a signal from which the pitch can be found using a standard estimator (e.g., one of those in [2]). Using a parallel approach, multi-channel pitch estimation [6] can be run in parallel with a broadband DOA estimator to obtain both pitch and DOA from the multi-channel signal. It is easy to see that, in multi-source scenarios, these estimation procedures may have problems with sources having the same DOA or overlapping harmonics.

As we have argued, joint DOA and pitch estimation (JDPE) methods are of interest as a robust alternative to cascaded or parallel approaches. Some approaches have recently been proposed, including the non-linear least squares (NLS) method [7], the spatio-temporal filtering based on the linearly constrained minimum variance (LCMV) beamformer [8], the correlation based method [9], and the subspace-based method [10]. While these methods perform well, they are computationally intensive, and faster methods may be required for some applications. More specifically, the methods of [7, 8] have cubic complexity for each combination of DOA and pitch candidates. Furthermore, none of these methods exploit that the pitch and DOA evolve in a piece-wise smooth manner.

In this paper, we present a fast algorithm for JDPE. The method is based on the frequency-domain minimum variance distortionless response (MVDR) beamformer, which is used to estimate the 2D spatio-temporal spectrum of the observed signal once per segment. From this 2D spectrum, the DOA and the pitch are estimated jointly by forming sums over the 2D spectrum for combinations of DOAs and pitches. This process essentially estimates the power of the assumed underlying periodic signal. Also, the number of harmonics is determined in the process using the maximum a posteriori method of [11, 12], something that is required to avoid ambiguities in the pitch estimates. Finally, the piece-wise smoothness of the DOA and pitch over time is exploited by the extension of the method [13] to include also the DOA.

The rest of this paper is organized as follows: in Section 2, we introduce the signal model and MVDR-based broadband beamforming, from which the proposed method is developed in Section 3. Later on, in Section 4, experimental results are reported. Finally, the paper is concluded in Section 5.

2 Problem Formulation

2.1 Signal Model

An array of M microphones receives broadband acoustic waves from D desired sound sources in a noisy environment without reverberation. We assume each desired complex signal, $s_d(n)$, is quasi-periodic with L_d number of harmonics, where $d = 1, ..., D$, and it is stationary over the sampling interval N. The signals captured by the mth microphone relating to the dth source are delayed by τ_{md} depending on their distance and the sampling frequency f_s. A linear combination of all sources besides an additive noise $e_m(n)$ constitutes this model:

$$x_m(n) = \sum_{d=1}^{D} s_d(n - f_s \tau_{md}) + e_m(n), \tag{E.1}$$

where

$$s_d(n - f_s\tau_{md}) = \sum_{l=1}^{L_d} a_{dl}\ e^{jl\omega_{0d}n}\, e^{-jl\omega_{0d}f_s\tau_{md}}, \tag{E.2}$$

and ω_{0d} is the fundamental frequency of the dth source. While many different array structures can be considered, we will assume a uniform linear array (ULA) structure herein for a proof of our concept. Consider a ULA denoted by M consecutive microphones with the specific inter-distance δ. Supposing a long distance from the ULA to the sources in comparison with δ, a plane wave and a homogeneous magnitude a_{dl} can be assumed across the array. By choosing the first microphone as a reference, the time delay between the other microphones and the reference is $\Delta\tau_{md} = (m-1)\delta\sin(\theta_d)/c$, where θ_d is the direction of desired signal, c is the wave propagation velocity and $\Delta\tau_{md} = \tau_{md} - \tau_{1d}$.

We can also formulate our signal model in the frequency domain, which is useful for deriving the proposed method. Stacking the spectral amplitudes of the observed signals at the M sensors for the frequency bin ω gives

$$\mathbf{X}(\omega) = [\, X_1(\omega)\, X_2(\omega)\, \ldots\, X_M(\omega)\,]^T. \tag{E.3}$$

By exploiting the relation between the sensor signals described by (E.2), the model further yields

$$\mathbf{X}(\omega) = \sum_{d=1}^{D} \mathbf{z}(\theta_d, \omega)\, S_d(\omega)\ +\ \mathbf{E}(\omega), \tag{E.4}$$

where $\mathbf{z}(\theta_d, \omega) = e^{-j\omega f_s\tau_{1d}}[\,1\ e^{-j\psi_{2d}}\ \ldots\ e^{-j\psi_{Md}}\,]^T$, and $\psi_{md} = \omega f_s \Delta\tau_{md}$.

2.2 MVDR Broadband Beamformer

The Capon beamformer, which is also known as the minimum variance distortionless response (MVDR) beamformer, is a type of baseband filter that can be extended to a broadband filter through a filter bank approach (FBA) [3]. To approach the proposed JDPE method, we introduce the broadband frequency-domain MVDR (FMV) algorithm as a quick solution in comparison with other time-domain beamformers [14].

First, a narrowband beamformer $\mathbf{W}(\theta, \omega)$ is designed to minimize the output power of the filter while it has a unit gain at a specific DOA, $\theta \in [-90^\circ, +90^\circ]$, and a sub-band frequency ω. In this way, we design a narrowband beamformer for a wide range of frequencies to get the broadband beamformer, i.e.,

$$\begin{aligned} &\min_{\mathbf{W}(\theta,\omega)} \mathbf{W}^H(\theta,\omega)\mathbf{R}_{\mathbf{X}}(\omega)\mathbf{W}(\theta,\omega) \\ &\text{s.t.} \quad \mathbf{W}^H(\theta,\omega)\mathbf{z}(\theta,\omega) = 1, \end{aligned} \tag{E.5}$$

where $\mathbf{R}_{\mathbf{X}}(\omega) \in \mathbb{C}^{MxM}$ is the correlation matrix of $\mathbf{X}(\omega)$ i.e., $\mathbf{R}_{\mathbf{X}}(\omega) = E\{\mathbf{X}(\omega)\mathbf{X}^H(\omega)\}$ that $E\{\cdot\}$ represents the expectation operation, and $\{\cdot\}^H$ represents the conjugate transpose of a matrix. The correlation matrix $\mathbf{R}_{\mathbf{X}}(\omega)$ is not known in most practical scenarios, so we estimate it as

$$\hat{\mathbf{R}}_{\mathbf{X}}(\omega) = \frac{1}{B}\sum_{b=1}^{B} \mathbf{X}_b(\omega)\,\mathbf{X}_b^H(\omega), \tag{E.6}$$

where $\mathbf{X}_b(\omega)$ denotes the bth complex spectral amplitude out of the last B estimates, and $\{\hat{\cdot}\}$ denotes the estimate. In practice, blocks of N samples are used to obtain the spectral amplitude estimates with consecutive blocks overlapping by Q samples.

The adaptive weights of the beamformer $\mathbf{W}(\theta,\omega)$ are formed using the Lagrange multiplier method [3] which yield

$$\mathbf{W}(\theta,\omega) = \frac{\hat{\mathbf{R}}_{\mathbf{X}}^{-1}(\omega)\mathbf{z}(\theta,\omega)}{\mathbf{z}^H(\theta,\omega)\hat{\mathbf{R}}_{\mathbf{X}}^{-1}(\omega)\mathbf{z}(\theta,\omega)}. \tag{E.7}$$

Inserting the optimal beamformer in the output power expression $\mathbf{W}^H(\theta,\omega)\hat{\mathbf{R}}_{\mathbf{X}}(\omega)\mathbf{W}(\theta,\omega)$ results in

$$J(\theta,\omega) = \frac{1}{\mathbf{z}^H(\theta,\omega)\hat{\mathbf{R}}_{\mathbf{X}}^{-1}(\omega)\mathbf{z}(\theta,\omega)}, \tag{E.8}$$

which should be an estimate of the spatio-temporal spectral power for θ and ω.

According to the designed filter, the estimated covariance matrix has to be invertible (E.7-E.8). This can be ensured by choosing $B \geq M$ in (E.6). In practice, B and Q should be chosen such that a good trade off between the robustness of the estimate and the validity of the stationary assumption is obtained.

3 Proposed Method

3.1 Order Estimation

To estimate the parameters (θ_d, ω_{0d}) of the desired signal source, we need an estimate of the number of harmonics L_d according to the signal model in (E.2). Here, we propose a model-order estimator, which is optimal in single source scenarios. The method is inspired by the maximum a posterior (MAP) estimator in [2, 11], where the noise variance is estimated using the directional spectrum obtained using the FMV method. It penalizes a maximum likelihood (ML) estimation method to find the maximum a posteriori probability of Φ and the number of harmonics, $L(\theta,\omega_0)$, given the temporal

spectrum at the candidate direction θ. In this way, we can estimate the model order for the relative pair of DOA and fundamental frequency [15]:

$$\hat{L}(\theta,\omega_0) = \arg\min_{L(\theta,\omega_0)} \{-\ln f(\mathbf{J}(\theta)|\Phi, L(\theta,\omega_0)) + \frac{1}{2}\ln|\hat{\mathbf{G}}|\}, \tag{E.9}$$

where $\mathbf{J}(\theta) = [\, J(\theta,0)\, J(\theta,\frac{2\pi}{N_f}) \ldots J(\theta,(\frac{N_f}{2}-1)\frac{2\pi}{N_f})\,]$, N_f is the length of the discrete Fourier transform (DFT), and $\mathbf{\Phi}$ denotes the vector containing the other estimation parameters: fundamental frequency, amplitudes, and phases. In the following, the notation of $L(\theta,\omega_0)$ is simplified to be exposed by L. The penalty part $\hat{\mathbf{G}}$ of this estimation is an approximation of the Fisher information matrix (FIM) relating to $\mathbf{\Phi}$, i.e.,

$$\hat{\mathbf{G}} \approx -E\left\{\frac{\partial^2 \ln f(\mathbf{J}(\theta)|\mathbf{\Phi}, L)}{\partial\mathbf{\Phi}\partial\mathbf{\Phi}^T}\right\}_{\mathbf{\Phi}=\hat{\mathbf{\Phi}}}. \tag{E.10}$$

The determinant of the given Hessian matrix $\hat{\mathbf{G}}$ can be normalized [15] with respect to the number of samples N (see [11]) as:

$$|\hat{\mathbf{G}}| = |\mathbf{K}^{-2}|\,|\mathbf{K}\,\hat{\mathbf{G}}\,\mathbf{K}|, \tag{E.11}$$

where it can be shown that $\mathbf{K}$ is given by

$$\mathbf{K} = \begin{bmatrix} N^{-\frac{3}{2}} & 0 \\ 0 & N^{-\frac{1}{2}}\mathbf{I}_{2Lx2L} \end{bmatrix}. \tag{E.12}$$

The estimate of the number of harmonics in (E.9) can be simplified by assuming that N is large, in which case [12] we obtain

$$\hat{L} \approx \arg\min_L \{N\ln\hat{\sigma}_L^2 + \frac{3}{2}\ln N + L\ln N\}, \tag{E.13}$$

where $\hat{\sigma}_L^2$ is the noise variance related to every candidates of a number of harmonics, and a fundamental frequency, ω_0. That is

$$\hat{\sigma}_L^2 = 2\left(\frac{N}{N_f}\sum_{i=0}^{N_f/2-1} J(\theta, i\frac{2\pi}{N_f}) - \sum_{l=1}^{L} J(\theta, l\omega_0) + r\right), \tag{E.14}$$

where, we have introduced the regularization factor r to account for inaccurate noise variance estimates for relatively small N.

3.2 Joint DOA and Pitch Estimation and Smoothing

A JDPE method for speech signals in a noisy field is proposed in this section based on the FMV method, and the general idea is depicted in Fig. E.1.

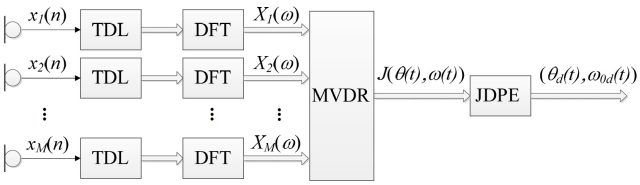


Fig. E.1: General diagram of the frequency-domain MVDR beamformer to joint DOA and pitch estimation (JDPE)

To estimate the fundamental frequency, the initial assumption of stationary temporal samples is introduced for an appropriate number of samples N in a tapped-delay line (TDL). The samples are mapped to the frequency domain using the DFT, and then the MVDR beamformer estimates a time-dependent 2D spectrum $J(\theta(t), \omega(t))$ at each time instance t.

According to the model in (E.2), the desired signals only have frequency contents at the harmonic frequencies. Hence, to estimate the pitch, we should only consider those bins. Therefore, we introduce the cost function

$$J_0(\theta_d(t), \omega_{0d}(t)) = \sum_{l=1}^{\hat{L}_d} J(\theta_d(t), l\omega_{0d}(t)). \tag{E.15}$$

Then, the pitch and the DOA are estimated jointly by maximizing $J_0(\theta_d(t), \omega_{0d}(t))$ for one source as

$$(\hat{\theta}_d(t), \hat{\omega}_{0d}(t)) = \arg \max_{(\theta_d(t), \omega_{0d}(t))} J_0(\theta_d(t), \omega_{0d}(t)). \tag{E.16}$$

The series of estimated parameters $[\hat{\theta}_d, \hat{\omega}_{0d}]$ have to be a continuous and smooth function of time according to the pitch and the position of the actual sound source. Smoothing of one dynamic parameter had been solved using a recursive algorithm in [13]. In this approach, a transition cost function $c(t, t_0)$ at time t is accumulated over the preceding states since t_0. The forward path is then the path that minimize the accumulated cost function $D(t)$ among all previous cost functions;

$$D(t) = \min_{t^*} \{D(t^*) + c(t^*, t_0)\} - B_{\mathrm{s}}, \tag{E.17}$$

where $t^* \in [t_0, t)$, B_{s} is a smoothing factor ($B_{\mathrm{s}} > 0$), and

$$c(t^*, t_0) = \frac{||[\hat{\theta}_d(t^*), \hat{\omega}_{0d}(t^*)] - [\hat{\theta}_d(t_0), \hat{\omega}_{0d}(t_0)]||_2}{(t^* - t_0)}. \tag{E.18}$$

Note that the transition cost function proposed here, is a generalization of the function in [13] from 1D to 2D.

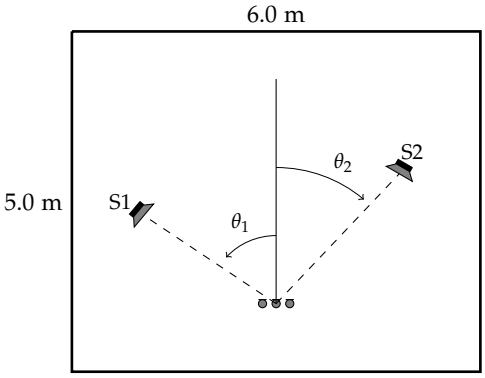


Fig. E.2: The room layout and the azimuth angles of sound sources; $\theta_1 \approx$ -53° and $\theta_2 \approx 45°$

4 Experimental Results

To evaluate the proposed JDPE method, we simulated a cocktail party in a room without reverberation with zero reflection order, as shown in Fig. E.2. The desired voiced speech, S1 uttering the sentence: "Why were you away a year Roy?", and a periodic signal with five harmonics, S2, as an interference are played simultaneously along with diffusive white noise. The sound propagation in a rectangular room is simulated using the image method [16]. This method simulates the room impulse response relating to the dimension, the reflection order and geometric position of acoustic sources and microphones [17]. We used this method to simulate acoustic waves on a ULA with three hyper-cardioid microphones, $M = 3$, at the specified positions with inter-distance of $\delta = 0.04$ m, and those were oriented respecting to the zero azimuth and elevation angles of the ULA. The distance between the microphones in the ULA should be smaller than half of the wavelength to avoid aliasing [3]. In addition, a real-life acoustic ambiance was simulated by adding diffusive acoustic noise, and the wave propagation speed was assumed $c = 343.2$ m/s. The spectrograms of the desired voiced speech and the interfered signal are shown in Fig. E.3. For the signal measured using the first microphone, the signal-to-interference-ratio (SIR) was 12.8 dB, while the signal-to-noise-ratio (SNR) was 10 dB.

Time-domain input signals were sampled across TDLs with $f_s = 8.0$ kHz of length $N = 256$, refreshed every 2.5 ms (20 time steps), and preserved in a buffer containing the $B = 10$ most recent ones. We calculated the DFTs of these vectors using zero-padded FFTs with a rectangular window of length $N_f = 4096$. The cross-correlation matrices for all frequency bins were then estimated from the B past DFTs, and they were used in three different methods: a parallel method, a cascade method, and the proposed method. The

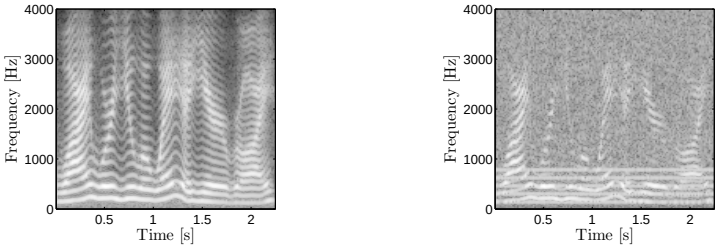


Fig. E.3: Spectrograms of the desired voiced speech (left), and the interfered mixture (right)

smoothing and the regularization factors in the proposed method were set to $B_s = 0.001\pi$ and $r = 10^{-6}$. In the parallel method, the FMV beamformer [14] runs along with the multi-channel pitch estimator in [6]. Finally, in the cascaded method, the NLS method [2] is applied after beamforming with the aforementioned parameters. Note that, in the cascaded method we used a Hanning window with 50% overlapping to recover the time domain signals.

Fig. E.4 depicts the results of the proposed JDPE method in comparison with the results of the cascaded and parallel methods. In order to evaluate the acquired results, we compared the estimates with the true DOA, $\theta_1 \approx$ -53°, and single channel pitch estimates of the clean signal obtained using the NLS method [2]. It shows the continuity and smoothness of both estimated DOA and pitch analytically, and the robust estimations are approved in terms of measured mean-square error (MSE) (see Table E.1).

5 Conclusion

In this paper, we have proposed the JDPE method. In this method, the DOA and pitch are estimated jointly by integrating the broadband MVDR spectrum over the harmonic frequencies for a set of candidate pitches and DOAs, and later on maximizing. The simplicity of the proposed method is a significant advantage in comparison with other joint estimation methods. The MVDR spectrum which can be implemented efficiently using FFTs, needs to be calculated once. It benefits the method as a fast spectral estimation. Our second contribution, is the spatio-temporal smoothing of the obtained

Table E.1: Mean square error (MSE) of estimated DOA and pitch of the different experiments

	MSE(ω_0) [Hz^2]	MSE(θ) [degree^2]
JDPE	122.1	1.5
Parallel	2.3x10^3	0.4
Cascade	2.9x10^3	0.4

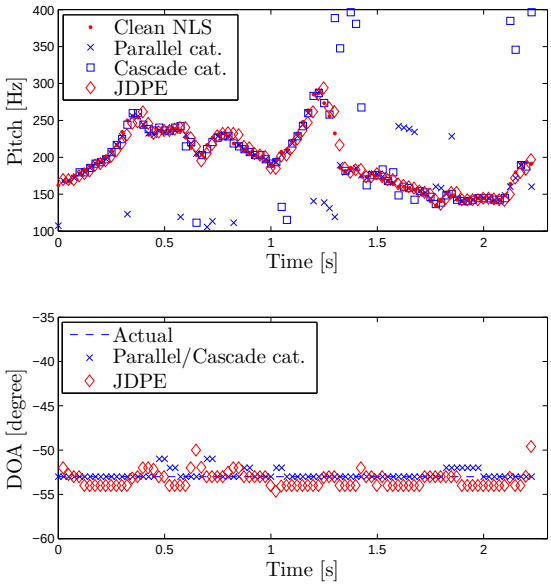


Fig. E.4: Estimation of pitch (top), and DOA (bottom) at the different experiments

DOA and pitch estimates using dynamic programming, which improves the robustness of the underlying estimator. Our simulations show that the proposed joint estimator outperforms traditional methods, i.e., a cascaded and a parallel approaches which estimate pitch and the DOA separately in a real-life scenario.

References

[1] J. Tabrikian, S. Dubnov, and Y. Dickalov, "Maximum a-posteriori probability pitch tracking in noisy environments using harmonic model," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 1, pp. 76 – 87, Jan. 2004.

[2] M. G. Christensen and A. Jakobsson, "Multi-pitch estimation," *Synthesis Lectures on Speech and Audio Process.*, vol. 5, no. 1, pp. 1–160, 2009.

[3] P. Stoica and R. Moses, *Spectral Analysis of Signals*. Pearson Education, Inc., 2005.

[4] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, Mar. 1986.

[5] Z. Zhou, M. G. Christensen, and H. C. So, "Two stage DOA and fundamental frequency estimation based on subspace techniques," in *Proc. IEEE Int. Conf. Signal Process.*, Oct. 2012.

[6] M. G. Christensen, "Multi-channel maximum likelihood pitch estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2012, pp. 409–412.

[7] J. R. Jensen, M. G. Christensen, and S. H. Jensen, "Nonlinear least squares methods for joint DOA and pitch estimation," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 21, no. 5, pp. 923 –933, May 2013.

[8] ——, "Joint DOA and fundamental frequency estimation methods based on 2-d filtering," in *Proc. European Signal Processing Conf.*, Aug. 2010, pp. 2091–2095.

[9] M. Wohlmayr and M. Képesi, "Joint position-pitch extraction from multichannel audio," in *Proc. Interspeech*, Aug. 2007, pp. 1629–1632.

[10] L. Y. Ngan, Y. Wu, H. C. So, P. C. Ching, and S. W. Lee, "Joint time delay and pitch estimation for speaker localization," in *Proc. IEEE Int. Symp. Circuits and Systems*, vol. 3, May 2003, pp. 722–725.

[11] P. M. Djuric, "Asymptotic MAP criteria for model selection," *IEEE Trans. Signal Process.*, vol. 46, no. 10, pp. 2726 –2735, Oct. 1998.

[12] P. Stoica and Y. Selen, "Model-order selection: a review of information criterion rules," *IEEE Signal Process. Mag.*, vol. 21, no. 4, pp. 36–47, Jul. 2004.

[13] H. Ney, "A dynamic programming algorithm for nonlinear smoothing," *Signal Processing*, vol. 5, no. 2, pp. 163–173, March 1983.

[14] M. E. Lockwood and et al., "Performance of time- and frequency-domain binaural beamformers based on recorded signals from real rooms," *J. Acoust. Soc. Am.*, vol. 115, no. 1, pp. 379–391, Jan. 2004.

[15] M. G. Christensen, J. L. Højvang, A. Jakobsson, and S. H. Jensen, "Joint fundamental frequency and order estimation using optimal filtering," *EURASIP J. on Applied Signal Process.*, vol. 2011, no. 1, pp. 1–18, Jun. 2011.

[16] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[17] E. A. P. Habets, “Room impulse response generator,” Technische Universiteit Eindhoven, Eindhoven, Netherlands, Tech. Rep., 2010, ver. 2.0.20100920.

Paper F

Fundamental Frequency and Model Order Estimation using Spatial Filtering

Sam Karimian-Azari, Jesper Rindom Jensen,
and Mads Græsbøll Christensen

The paper has been published in the
Proceeding IEEE Int. Conf. Acoust., Speech, Signal Process., pp. 5964–5968, 2014.

The layout has been revised.

Abstract

In signal processing applications of harmonic-structured signals, estimates of the fundamental frequency and number of harmonics are often necessary. In real scenarios, a desired signal is contaminated by different levels of noise and interferers, which complicate the estimation of the signal parameters. In this paper, we present an estimation procedure for harmonic-structured signals in situations with strong interference using spatial filtering, or beamforming. We jointly estimate the fundamental frequency and the constrained model order through the output of the beamformers. Besides that, we extend this procedure to account for inharmonicity using unconstrained model order estimation. The simulations show that beamforming improves the performance of the joint estimates of fundamental frequency and the number of harmonics in low signal to interference (SIR) levels, and an experiment on a trumpet signal show the applicability on real signals.

1 Introduction

In real life, we often have multiple signal sources present at the same time, which has a detrimental impact on the quality and intelligibility of a recorded speech signal. We can improve the quality of a desired signal by choosing an appropriate enhancement method, which can be categorized in three different groups: statistical, filtering, and subspace methods [1]. In the enhancement of harmonic-structured signals as considered here, e.g., voiced speech, both the fundamental frequency and number of harmonics estimates are necessary in filter designs (for example [2–4]). Therefore, we require to estimate these parameters. The estimation of the fundamental frequency, or pitch in audio signal processing, is a challenging problem with applications in enhancement, separation, classification, compression, etc., and different methods have been investigated in the single-channel case [1, 5]. The estimation of number of harmonics is another problem in enhancement of harmonic-structured signals. This integer-valued parameter relating to the number of sinusoidal components must be estimated from the received signals to yield accurate pitch estimates and high-quality enhancement, and some methods have been investigated in the single-channel case [6].

In most of the state-of-the-art methods for fundamental frequency and number of harmonics estimations, the desired signal is assumed to be degraded by additive white Gaussian noise [7–10]. For example, the Markov-like weighted least-squares (WLS) [11] (see also [1, 12]) and the maximum a posteriori (MAP) [6, 13] methods are fundamental frequency and number of harmonics estimators for only one signal source. In a situation with the presence of interference having the harmonic structure, which is very common, some methods are available to estimate the parameters of multiple signal

sources [1]. In these methods, the basic assumption is that the desired signal has higher power than the interferers [2, 14], something that is not always the case. Besides that, multiple harmonic-structured signals with spectral overlap may cause a wrong estimate of the fundamental frequency and the number of harmonics. Furthermore, the inharmonicity problem [15], which is the phenomenon that the frequencies of the harmonics are not exact integers of a fundamental frequency, results in a model mismatch, and leads to biased parameter estimates, e.g., in stiff-stringed instruments.

Exploiting spatial separation is a solution to separate multiple signals using multiple microphones, and beamforming is one such technique to estimate the signal arriving from the desired direction [16] using different source localization methods which have been investigated in [17]. In this paper, we estimate both the fundamental frequency and the number of harmonics, which we call the model order, of a harmonic-structured signal using a beamforming technique to separate the desired signal from high power interferers, which are spatially separated, e.g., by using broadband minimum variance distortionless response (MVDR) [18, 19] beamforming. We can also estimate the model order of the desired signal from the output of the beamformer [20] using the MAP method with the constrained harmonic-model, consisting of a fundamental frequency and its integers. Because of the problem of inharmonicity and harmonic frequencies mismatch, we extend this method for the unconstrained model, consisting of independent sinusoidal components, to estimate both the fundamental frequency and model order. Then the fundamental frequency estimate will be performed using the WLS method [11].

The rest of this paper is organized as follows. In Section 2, we introduce the multi-source signal model that the work is based on and apply it in beamforming. In Section 3, we derive the constrained and unconstrained model order and fundamental frequency estimates, and then explore the results of simulations in Section 4. In closing, the work is discussed in Section 5 along with its relation to state-of-the-art.

2 Problem Formulation

2.1 Signal Model

We consider N independent sources of harmonic acoustic waves, which are placed at different spatial positions, that propagate acoustic waves from their respective direction of arrival (DOA), i.e., θ_n for $n = 1, \ldots, N$, relative to a receiver. We assume a microphone array with a set of M omnidirectional microphones receives these acoustic waves besides random noise, i.e., $y_m(t)$ and $v_m(t)$ for $m = 1, \ldots, M$. Then, we model the combination of harmonic-structured signal sources, i.e., $x_n(t) = \sum_{l=1}^{L_n} \alpha_{n,l}\, e^{\jmath(l\omega_n t + \varphi_{n,l})}$ that ω_n is the

fundamental frequency with L_n number of harmonics with the magnitude of $\alpha_{n,l}$ and phase of $\varphi_{n,l}$,

$$y_m(t) = \sum_{n=1}^{N} \sum_{l=1}^{L_n} \alpha_{n,l}\, e^{\jmath(l\omega_n t + \varphi_{n,l})} e^{-\jmath l\omega_n \Delta\tau_{m,n}} + v_m(t), \tag{F.1}$$

where $\jmath = \sqrt{-1}$, and $\Delta\tau_{m,n}$ is the time difference of arrival between the mth and the first microphone for the nth source. By expressing the signal model (F.1) in the frequency-domain vector notation [19], the received broadband signals $\mathbf{Y}(\omega) = [\, Y_1(\omega) \dots\ Y_M(\omega)\,]^{\mathrm{T}}$ are formulated as functions of the steering vector $\mathbf{d}(\theta_n, \omega)$, signal sources $X_n(\omega)$, and noise $\mathbf{V}(\omega)$, defined similar to $\mathbf{Y}(\omega)$, as

$$\mathbf{Y}(\omega) = \sum_{n=1}^{N} \mathbf{d}(\theta_n, \omega) X_n(\omega) + \mathbf{V}(\omega), \tag{F.2}$$

where the steering vector is the set of phase shifts between microphones defined at each subband by choosing the first microphone as the reference

$$\mathbf{d}(\theta_n, \omega) = [\, 1 \;\; e^{-\jmath\omega\Delta\tau_{2,n}} \;\; \dots \;\; e^{-\jmath\omega\Delta\tau_{M,n}} \,]^{\mathrm{T}}. \tag{F.3}$$

With the aim of the spatial source separation, we can write the spatial correlation matrix, by the assumption of uncorrelated signal sources and noise, as

$$\begin{aligned} \mathbf{R}_{\mathbf{Y}}(\omega) &= \mathrm{E}\{\mathbf{Y}(\omega)\mathbf{Y}^{\mathrm{H}}(\omega)\} \\ &= \sum_{n=1}^{N} \mathbf{d}(\theta_n, \omega) J_{X_n}(\omega) \mathbf{d}^{\mathrm{H}}(\theta_n, \omega) + \mathbf{R}_{\mathbf{V}}(\omega), \end{aligned} \tag{F.4}$$

where $\mathrm{E}\{\cdot\}$ denotes mathematical expectation, and the superscript $^{\mathrm{H}}$ the transpose-conjugate operator. We define $J_{X_n}(\omega) = \mathrm{E}\{|X_n(\omega)|^2\}$ as the subband power of each signal source, and the noise correlation matrix as $\mathbf{R}_{\mathbf{V}}(\omega) = \mathrm{E}\{\mathbf{V}(\omega)\mathbf{V}^{\mathrm{H}}(\omega)\}$.

2.2 Spatial Filtering

All the complex values of the microphone outputs at the subband ω are applied to a complex-valued spatial filter $\mathbf{H}(\theta, \omega)$, or a beamformer as we refer to it, of the length M at each candidate direction θ subject to $\mathbf{H}^{\mathrm{H}}(\theta, \omega)\mathbf{d}(\theta, \omega) = 1$. In general, the output signal will be

$$Z(\theta, \omega) = \mathbf{H}^{\mathrm{H}}(\theta, \omega)\mathbf{Y}(\omega), \tag{F.5}$$

and the output power of the designed filters is

$$\begin{aligned} J_Z(\theta,\omega) &= \mathrm{E}\{Z(\theta,\omega)Z^{\mathrm{H}}(\theta,\omega)\} \\ &= \mathbf{H}^{\mathrm{H}}(\theta,\omega)\mathbf{R}_{\mathbf{Y}}(\omega)\mathbf{H}(\theta,\omega). \end{aligned} \tag{F.6}$$

By considering $X_1(\omega)$ as the desired signal, and substituting (F.4) into (F.6) at the direction of the desired signal, i.e., θ_1, we acquire the output power of the beamformer as

$$\begin{aligned} J_Z(\theta_1,\omega) =\ & J_{X_1}(\omega) + \mathbf{H}^{\mathrm{H}}(\theta_1,\omega)\mathbf{R}_{\mathbf{V}}(\omega)\mathbf{H}(\theta_1,\omega) \\ & + \sum_{n=2}^{N} \mathbf{H}^{\mathrm{H}}(\theta_1,\omega)\mathbf{d}(\theta_n,\omega)J_{X_n}(\omega)\mathbf{d}^{\mathrm{H}}(\theta_n,\omega)\mathbf{H}(\theta_1,\omega) \\ =\ & J_{X_1}(\omega) + \Psi(\theta_1,\omega), \end{aligned} \tag{F.7}$$

where $\Psi(\theta_1,\omega)$ is a residual noise-plus-interference after filtering. The broadband output power of the filter, and the broadband output power of the noise-plus-interference are, respectively,

$$J_Z(\theta) = \frac{1}{2\pi}\int_0^{2\pi} J_Z(\theta,\omega)\mathrm{d}\omega, \tag{F.8}$$

$$\Psi(\theta_1) = \frac{1}{2\pi}\int_0^{2\pi} \Psi(\theta_1,\omega)\mathrm{d}\omega = J_Z(\theta_1) - J_{X_1}, \tag{F.9}$$

where J_{X_1} is the broadband power of the desired signal.

The delay-and-sum (DS) beamformer is designed based on the principle that the directivity pattern of the filter is steered to the DOA of interest, i.e., $\mathbf{H}_{\mathrm{DS}}(\theta,\omega) = \mathbf{d}(\theta,\omega)/M$, and the desired signal can be filtered in the composition of different signals (F.2) depending on the respective DOA. Besides the directivity pattern criteria, the minimum variance distortionless response (MVDR) beamformer is designed to minimize the output power

$$\begin{aligned} &\min_{\mathbf{H}(\theta,\omega)} \mathbf{H}^{\mathrm{H}}(\theta,\omega)\mathbf{R}_{\mathbf{Y}}(\omega)\mathbf{H}(\theta,\omega) \\ &\text{s.t.} \quad \mathbf{H}^{\mathrm{H}}(\theta,\omega)\mathbf{d}(\theta,\omega) = 1, \end{aligned} \tag{F.10}$$

then the optimal MVDR filter is given by [21]

$$\mathbf{H}_{\mathrm{MVDR}}(\theta,\omega) = \frac{\mathbf{R}_{\mathbf{Y}}^{-1}(\omega)\mathbf{d}(\theta,\omega)}{\mathbf{d}^{\mathrm{H}}(\theta,\omega)\mathbf{R}_{\mathbf{Y}}^{-1}(\omega)\mathbf{d}(\theta,\omega)}. \tag{F.11}$$

3 Proposed Method

The signal source X_n with an integer number of harmonics, i.e., L_n, can be modeled in two ways: the constrained (C) harmonic-model that is the integration of integer frequency coefficients relating to the fundamental frequency

ω_n, i.e.,

$$\mathbb{X}_n^{\mathrm{C}}(\omega_n) = [\ X_n(\omega_n)\ X_n(2\omega_n) \ldots\ X_n(L_n^{\mathrm{C}}\omega_n)\ ]^{\mathrm{T}}, \tag{F.12}$$

and the unconstrained (UC) model that is the integer number of independent periodic signals, i.e.,

$$\mathbb{X}_n^{\mathrm{UC}}(\mathbf{\Omega}_n) = [\ X_n(\omega_{n,1})\ X_n(\omega_{n,2}) \ldots\ X_n(\omega_{n,L_n^{\mathrm{UC}}})\ ]^{\mathrm{T}}, \tag{F.13}$$

where $\mathbf{\Omega}_n = [\ \omega_{n,1}\ \omega_{n,2}\ \ldots\ \omega_{n,L_n^{\mathrm{UC}}}\ ]^{\mathrm{T}}$ is a set of unconstrained frequencies. By the assumption of two models, the power of the desired signal can be estimated as

$$J_{X_1}^{\mathrm{C}}(\omega_1) = 2 \parallel \mathbb{X}_1^{\mathrm{C}}(\omega_1) \parallel_2^2, \tag{F.14}$$

$$J_{X_1}^{\mathrm{UC}}(\mathbf{\Omega}_1) = 2 \parallel \mathbb{X}_1^{\mathrm{UC}}(\mathbf{\Omega}_1) \parallel_2^2. \tag{F.15}$$

We can estimate the model order of a harmonic signal from the output power of a beamformer at the desired direction by minimizing the broadband noise power [20]. For both the constrained and unconstrained models in (F.12) and (F.13), we write the broadband output power of the noise-plus-interference from (F.9) like

$$\Psi^{\mathrm{C}}(\theta_1) = J_Z(\theta_1) - J_{X_1}^{\mathrm{C}}(\omega_1), \tag{F.16}$$

$$\Psi^{\mathrm{UC}}(\theta_1) = J_Z(\theta_1) - J_{X_1}^{\mathrm{UC}}(\mathbf{\Omega}_1). \tag{F.17}$$

With the assumption of white Gaussian noise and using N_f frequency samples, we can jointly estimate the fundamental frequency and the number of constrained harmonics using the MAP method in the model order estimation [1, 6] as

$$(\hat{L}_1^{\mathrm{C}}, \hat{\omega}_1^{\mathrm{C}}) \approx \arg \min_{L_1^{\mathrm{C}}, \omega_1} \{ N_f \ln[\Psi^{\mathrm{C}}(\theta_1)] + \frac{3}{2} \ln N_f + L_1^{\mathrm{C}} \ln N_f \}, \tag{F.18}$$

which consists of the log-likelihood function of the noise-plus-interference and the penalty part. The penalty part is estimated through the normalization of the Fisher information matrix for a candidate fundamental frequency and L_1 related amplitudes and phases [14]. We can extend this method for estimating the number of independent sinusoids and the related amplitudes and phases, i.e.,

$$(\hat{L}_1^{\mathrm{UC}}, \hat{\mathbf{\Omega}}_1) \approx \arg \min_{L_1^{\mathrm{UC}}, \mathbf{\Omega}_1} \{ N_f \ln[\Psi^{\mathrm{UC}}(\theta_1)] + \frac{5}{2} L_1^{\mathrm{UC}} \ln N_f \}. \tag{F.19}$$

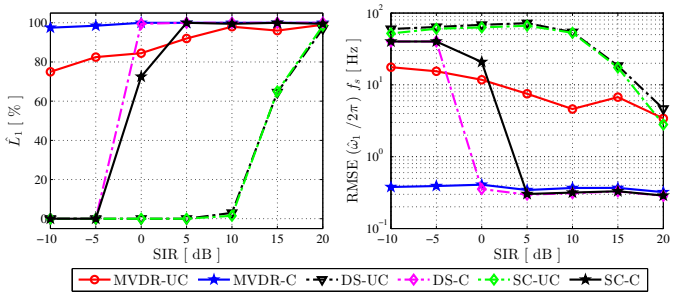


Fig. F.1: Performance of the model order and the fundamental frequency estimators for different SIRs [dB], with SNR = 20 dB, and $\Delta\omega_n/2\pi$ = 0.00025.

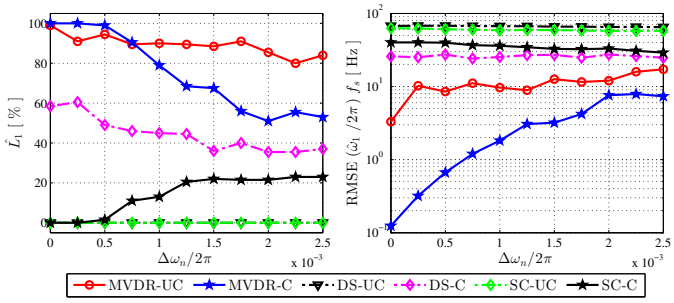


Fig. F.2: Performance of the model order and the fundamental frequency estimators for different $\Delta\omega_n/2\pi$, with SNR = 20 dB, and SIR = -1.5 dB.

To estimate the fundamental frequency that has the best match to the frequency estimates obtained using the unconstrained model, i.e., $\hat{\mathbf{\Omega}}_1$, we apply the WLS method [11]:

$$\hat{\omega}_1^{\mathrm{UC}} = \frac{\sum_{l=1}^{\hat{L}^{\mathrm{UC}}} l\ |X_1(\omega_{1,l})|^2\ \omega_{1,l}}{\sum_{l=1}^{\hat{L}^{\mathrm{UC}}} l^2\ |X_1(\omega_{1,l})|^2}. \tag{F.20}$$

4 Simulation Results

In the following, we evaluate the proposed method and compare the results with single-channel (SC) results in different experiments using synthetic data, and also in a simulation with a real trumpet sound. Then, we measure the root mean square errors (RMSEs) of the fundamental frequency and percentage of correctly model order estimates from 200 Monte-Carlo simulations. In all simulations, we place two synthetic signals at $\theta_1 = 60^\circ$ and $\theta_2 = 40^\circ$, where $\omega_1/2\pi = 0.0225$, $L_1 = 5$ with unit amplitudes, and $\omega_2/2\pi = 0.0275$, $L_2 = 7$, with equal amplitudes depending on signal to inter-

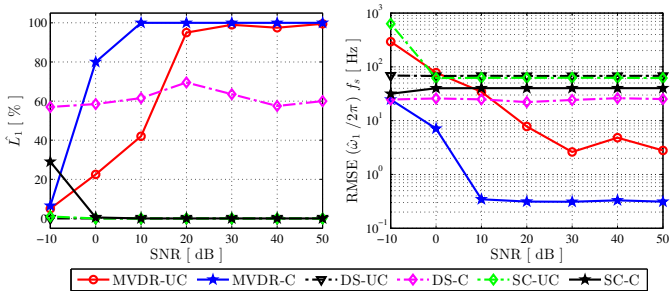


Fig. F.3: Performance of the model order and the fundamental frequency estimators for different SNRs [dB], with $\Delta\omega_n/2\pi$ = 0.00025, and SIR = -1.5 dB.

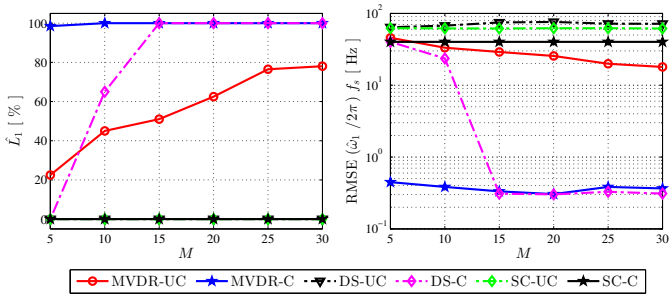


Fig. F.4: Performance of the model order and the fundamental frequency estimators using different number of microphones, with $\Delta\omega_n/2\pi$ = 0.00025, SNR = 10 dB, and SIR = -1.5 dB.

ference ratio (SIR) levels, and the sampling frequency is $f_s = 8.0$ kHz. These harmonic-structured signals are simulated like $\mathbf{\Omega}_1 = [\ (\omega_1 + \Delta\omega_{1,1})\ (2\omega_1 + \Delta\omega_{1,2})\ \ldots\ (L_1\omega_n + \Delta\omega_{1,L_1})\ ]^{\mathrm{T}}$, where the $\Delta\omega_{1,l}$ is a normal distribution of the frequencies with a variance of zero for simulating the constrained harmonic-model, and a non-zero variance for the unconstrained model with perturbed harmonics.

We model a uniform linear array (ULA) consisting of $M = 10$ omnidirectional microphones, for which the distance between two successive sensors is $\delta = 0.04$ m (smaller than half of the minimum wavelength $\delta \leq \lambda_{\min}/2$), and add independent white Gaussian noise to each microphone depending on signal to noise ratio (SNR) levels. The time differences of arrival is $\Delta\tau_{m,n} = (m-1)\delta\sin(\theta_n)/\mathrm{c}$, where the wave propagation speed is assumed to be $c = 343.2$ m/s. The mathematical expectation is estimated by time averaging of B temporal frames [20, 22]. In the MVDR beamforming design (F.11), the full rank correlation matrix can be guaranteed by choosing $B \geq M$, so that, we choose $B = 30$ in all simulations.

First, the spectral amplitudes of each subband are estimated using a 512 point discrete Fourier transform (DFT). Then, for spectral estimation with large frequency grids, the 65536 point DFT is taken from the zero-padded

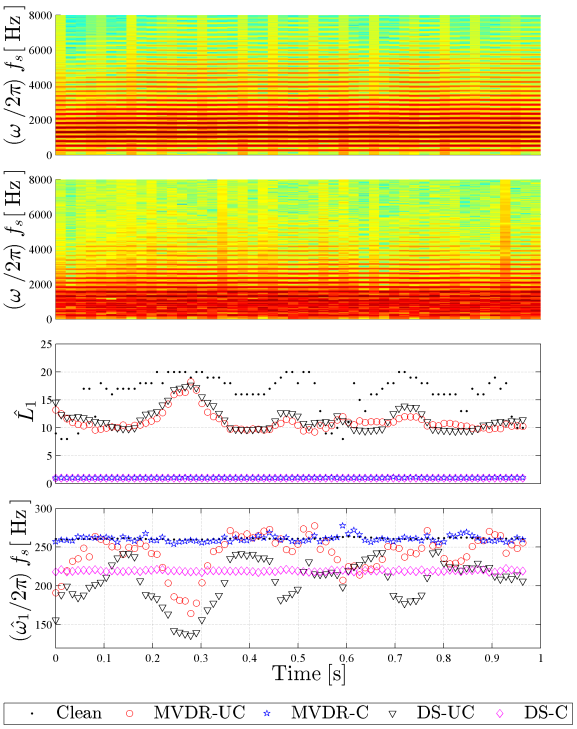


Fig. F.5: According to the order of plots from top to down: the spectrograms of a clean trumpet signal $|X_1(\omega)|$ and the distorted signal $|Y_1(\omega)|$, and the estimates of the number of harmonics and the fundamental frequency.

inverse-DFT of the output signal from the beamformers, and the broadband output power in (F.18) and (F.19) are normalized like in [20]. Figure F.1 shows that the fundamental frequency and the model order estimation methods are performed in low SIRs using beamforming, and the MVDR beamformer performs better than the DS beamformer. Figure F.2 indicates that the unconstrained model order estimate is more accurate in comparison with the constrained harmonic-model in high ranges of perturbed harmonics, $\Delta\omega_n/2\pi \geq 0.001$. The MVDR beamformer outperforms the DS beamformer in low SNRs and number of microphones in figures F.3 and G.2, respectively. We also conduct an experiment on a trumpet signal with vibrato, as the desired signal, which is corrupted by a synthetic signal similar to in the previous simulations and white Gaussian noise, i.e., SIR = -1.5 dB and SNR = 10 dB. Figure F.5 indicates that the unconstrained model order has better estimates than the other model, and the fundamental frequency estimates via the constrained model has better results, compared with the clean signal estimates using the constrained harmonic-model.

5 Discussion and Conclusion

In this paper, we improve the fundamental frequency and model order estimates of harmonic-structured signals in situations with low SIR. In the multi-channel parameter estimation methods in [10] and [8], it has been considered that a desired signal is contaminated only by Gaussian noise, although in situations with spatially separated interference signals, which are likely in real scenarios, the joint fundamental frequency and constrained model order estimates [14] can be facilitated using spatial filters [20]. Simulations show beamforming will yield better results than the corresponding single-channel estimates, and the optimal MVDR beamformer outperforms the DS, as an example, for closely spaced signal sources. Moreover, through the MAP model order estimation with a uniform probability distribution of random candidates, a general unconstrained model is approached instead of a particular model in [15]. To approach high-resolution of spectral estimates with a minimum variance capability, the DFT method, which we used in our experiments, can be replaced by different methods [5], e.g., unconstrained model extension of the methods in [14] and [23], note that also in the two-dimensional MVDR filter design [23].

References

[1] M. G. Christensen and A. Jakobsson, "Multi-pitch estimation," *Synthesis Lectures on Speech and Audio Process.*, vol. 5, no. 1, pp. 1–160, 2009.

[2] ——, "Optimal filter designs for separating and enhancing periodic signals," *IEEE Trans. Signal Process.*, vol. 58, no. 12, pp. 5969–5983, Dec. 2010.

[3] A. Nehorai and B. Porat, "Adaptive comb filtering for harmonic signal enhancement," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 5, pp. 1124–1138, Oct. 1986.

[4] W. Jin, X. Liu, M. Scordilis, and L. Han, "Speech enhancement using harmonic emphasis and adaptive comb filtering," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 18, no. 2, pp. 356–368, Feb 2010.

[5] P. Stoica and R. Moses, *Spectral Analysis of Signals*. Pearson Education, Inc., 2005.

[6] P. Stoica and Y. Selen, "Model-order selection: a review of information criterion rules," *IEEE Signal Process. Mag.*, vol. 21, no. 4, pp. 36–47, Jul. 2004.

[7] J. Tabrikian, S. Dubnov, and Y. Dickalov, "Maximum a-posteriori probability pitch tracking in noisy environments using harmonic model," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 1, pp. 76 – 87, Jan. 2004.

[8] Z. Zhou, H. So, and M. Christensen, "Parametric modeling for damped sinusoids from multiple channels," *IEEE Trans. Signal Process.*, vol. 61, no. 15, pp. 3895–3907, Aug 2013.

[9] J. Benesty, J. Chen, Y. Huang, and I. Cohen, *Noise Reduction in Speech Processing*. Springer-Verlag, 2009.

[10] M. G. Christensen, "Multi-channel maximum likelihood pitch estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2012, pp. 409–412.

[11] H. Li, P. Stoica, and J. Li, "Computationally efficient parameter estimation for harmonic sinusoidal signals," *Elsevier Signal Process.*, vol. 80(9), pp. 1937–1944, 2000.

[12] M. G. Christensen, P. Vera-Candeas, S. D. Somasundaram, and A. Jakobsson, "Robust subspace-based fundamental frequency estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2008, pp. 101–104.

[13] P. M. Djuric, "Asymptotic MAP criteria for model selection," *IEEE Trans. Signal Process.*, vol. 46, no. 10, pp. 2726 –2735, Oct. 1998.

[14] M. G. Christensen, J. L. Højvang, A. Jakobsson, and S. H. Jensen, "Joint fundamental frequency and order estimation using optimal filtering," *EURASIP J. on Applied Signal Process.*, vol. 2011, no. 1, pp. 1–18, Jun. 2011.

[15] T. D. Rossing, F. R. Moore, and P. A. Wheeler, *The Science of Sound*, 3rd ed. Addison Wesley, 2002.

[16] B. D. Van Veen and K. M. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE ASSP Mag.*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[17] M. S. Brandstein and H. Silverman, "A practical methodology for speech source localization with microphone arrays," *Comput. Speech Language*, 1997.

[18] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.

[19] J. Benesty, J. Chen, and E. A. P. Habets, *Speech Enhancement in the STFT Domain*. Springer, 2012, vol. 5.

[20] S. Karimian-Azari, J. R. Jensen, and M. G. Christensen, "Fast joint DOA and pitch estimation using a broadband MVDR beamformer," in *Proc. European Signal Processing Conf.*, Sept. 2013, pp. 1–5.

[21] J. Benesty, Y. Huang, and J. Chen, *Microphone Array Signal Processing*. Springer-Verlag, 2008, vol. 1.

[22] M. E. Lockwood and et al., "Performance of time- and frequency-domain binaural beamformers based on recorded signals from real rooms," *J. Acoust. Soc. Am.*, vol. 115, no. 1, pp. 379–391, Jan. 2004.

[23] A. Jakobsson, S. L. Jr. Marple, and P. Stoica, "Computationally efficient two-dimensional Capon spectrum analysis," *IEEE Trans. Signal Process.*, vol. 48, no. 9, pp. 2651–2661, Sep. 2000.

Paper G

A Broadband Beamformer using Controllable Constraints and Minimum Variance

Sam Karimian-Azari, Jacob Benesty, Jesper Rindom Jensen, and Mads Græsbøll Christensen

The paper has been published in the
Proceeding European Signal Processing Conf., pp. 666–670, 2014.

The layout has been revised.

Abstract

The minimum variance distortionless response (MVDR) and the linearly constrained minimum variance (LCMV) beamformers are two optimal approaches in the sense of noise reduction. The LCMV beamformer can also reject interferers using linear constraints at the expense of reducing the degree of freedom in a limited number of microphones. However, it may magnify noise that causes a lower output signal-to-noise ratio (SNR) than the MVDR beamformer. Contrarily, the MVDR beamformer suffers from interference in output. In this paper, we propose a controllable LCMV (C-LCMV) beamformer based on the principles of both the MVDR and LCMV beamformers. The C-LCMV approach can control a compromise between noise reduction and interference rejection. Simulation results show that the C-LCMV beamformer outperforms the MVDR beamformer in interference rejection, and the LCMV beamformer in background noise reduction.

1 Introduction

Multiple acoustic sources are usually present in real situations. For speech processing applications such as teleconferencing and hearing aids, noise reduction techniques are developed to achieve a high quality and preserve the intelligibility of the desired signal. In single-channel signal enhancement methods, both the desired signal and noise are filtered at the same time [1]. While the speech quality is increased in the Wiener filter, which is an example of a known noise-reduction filter [1, 2], speech distortion will be increased in the presence of interference. Exploiting spatial separation is another solution to separate multiple signals and enhance the desired signal using multiple microphones, which is called a microphone array.

Beamforming is one of the techniques for microphone arrays to estimate the signal arriving from a desired direction-of-arrival (DOA), and separate different signal sources [3]. The basic principle is that the received signals through multiple microphones are synchronized by delays depending on the desired DOA using complex weighted filters and summed, e.g., as in the delay-and-sum (DS) beamforming [4]. Besides the spatial separation, signal enhancement is another issue in the filter design, where the adaptive filters are designed to minimize the noise and interference using the statistics of the received signals. An adaptive multichannel filter can provide a trade-off between noise reduction and signal distortion [5], e.g., the multichannel Wiener filter [6], and the maximum SNR filter [4]. Some well-known examples of beamformer designs are the least-squares, multiple sidelobe canceler (MSC) [7], generalized sidelobe canceler (GSC) [8, 9], superdirective [10], minimum variance distortionless response (MVDR) [11], and linearly constrained minimum variance (LCMV) [12] beamformers. For more details about vari-

ous beamformer designs, we refer the reader to [4] and [13].

In this paper, we propose a new beamformer based on the principles of the MVDR and LCMV beamformers using the spectral decomposition [14–16]. Both are designed to minimize the output power subject to a unit output gain at the desired DOA, and through exploiting the decomposition of interfering signals, we can have multiple constraints to reject the interference in the LCMV. Although the number of constraints degrades degree of freedom (DOF) of a beamformer which is as many as the number of microphones [17]. Though there is a trade-off between noise and interference reduction, the LCMV beamformer may magnify the background noise [18] with having high sidelobes [19]. Therefore, we explore a new flexible beamformer based on the paradigm of minimum variance in order to control the output signal-to-interference-plus-noise ratio (SINR) and signal-to-interference ratio (SIR). That is, we propose the controllable LCMV (C-LCMV) beamformer with a variable number of constraints.

The rest of this paper is organized as follows. In Section 2, we model the composition of multiple signal sources in vector notation and design the MVDR and LCMV beamformers accordingly. In Section 3, we propose the C-LCMV beamformer, and then explore the properties of this method in simulations in Section 4. The work is concluded in Section 5.

2 Problem Formulation

2.1 Signal Model

We consider a microphone array, consisting of M omnidirectional microphones, receives broadband signals from N acoustic sources besides a background noise, where $N \leq M$. In general, we model the received signals at the frequency index f in a vector notation as $\mathbf{y}(f) = [\, Y_1(f)\ Y_2(f)\ \cdots\ Y_M(f)\,]^{\mathrm{T}}$, where $Y_m(f)$ is the mth microphone narrowband signal and the superscript $^{\mathrm{T}}$ is the transpose operator. We write the vector $\mathbf{y}(f)$ as a function of the (known) steering vectors $\mathbf{d}_n(f)$ and signal sources $X_n(f)$ for $n = 1, ..., N$ [4, 16] like

$$\begin{aligned}\mathbf{y}(f) &= \mathbf{d}_1(f)X_1(f) + \sum_{n=2}^{N} \mathbf{d}_n(f)X_n(f) + \mathbf{v}(f) \\ &= \mathbf{D}(f)\mathbf{x}(f) + \mathbf{v}(f),\end{aligned} \tag{G.1}$$

where $\mathbf{v}(f) = [\, V_1(f)\ V_2(f)\ \cdots\ V_M(f)\,]^{\mathrm{T}}$ is the additive background noise, $\mathbf{x}(f) = [\, X_1(f)\ X_2(f)\ \cdots\ X_N(f)\,]^{\mathrm{T}}$ is the collection of signal sources, and we define $\mathbf{D}(f)$ as the $M \times N$ matrix containing all steering vectors relating to

the N signal sources, i.e.,

$$\mathbf{D}(f) = [\, \mathbf{d}_1(f)\ \mathbf{d}_2(f)\ \cdots\ \mathbf{d}_N(f)\,]. \tag{G.2}$$

We assume that $X_n(f)$ and $V_m(f)$ are uncorrelated and zero mean. Furthermore, we consider $X_1(f)$ as the desired signal that we wish to extract from the observations, while $X_n(f)$ for $n = 2, 3, \ldots, N$ are interferers.

The correlation matrix of $\mathbf{y}(f)$ is defined as $\mathbf{\Phi}_{\mathbf{y}}(f) = E\left[\mathbf{y}(f)\mathbf{y}^{\mathrm{H}}(f)\right]$, where $E[\cdot]$ denotes mathematical expectation, and the superscript $^{\mathrm{H}}$ is the transpose-conjugate operator. If we assume all signal sources and noise are uncorrelated, we can write the correlation matrix as

$$\begin{aligned}\mathbf{\Phi}_{\mathbf{y}}(f) &= \mathbf{D}(f)\,\mathbf{\Phi}_{\mathbf{x}}(f)\,\mathbf{D}^{\mathrm{H}}(f) + \mathbf{\Phi}_{\mathbf{v}}(f) \\ &= \mathbf{d}_1(f)\,\phi_{X_1}(f)\,\mathbf{d}_1^{\mathrm{H}}(f) + \mathbf{\Phi}_{\mathrm{in}}(f) + \mathbf{\Phi}_{\mathbf{v}}(f),\end{aligned} \tag{G.3}$$

where $\mathbf{\Phi}_{\mathbf{x}}(f) = \mathrm{diag}[\,\phi_{X_1}(f)\ \phi_{X_2}(f)\ \ldots\ \phi_{X_N}(f)\,]$ is a diagonal matrix of size $N \times N$ containing the variances of the sources at the frequency index f, i.e., $\phi_{X_n}(f) = E\left[\,|X_n(f)|^2\,\right]$, the correlation matrix of $\mathbf{v}(f)$ is $\mathbf{\Phi}_{\mathbf{v}}(f) = E\left[\mathbf{v}(f)\mathbf{v}^{\mathrm{H}}(f)\right]$, and $\mathbf{\Phi}_{\mathrm{in}}(f) = \sum_{n=2}^{N} \mathbf{d}_n(f)\,\phi_{X_n}(f)\,\mathbf{d}_n^{\mathrm{H}}(f)$ is the interference correlation matrix. If the components of the steering vectors are only phase shifts, which is usually the case, then $\mathbf{d}_n^{\mathrm{H}}(f)\mathbf{d}_n(f) = M$. As a result, we can deduce the narrowband input SIR and input SINR respectively like

$$\mathrm{iSIR}(f) = \frac{\phi_{X_1}(f)}{\sum_{n=2}^{N} \phi_{X_n}(f)}, \tag{G.4}$$

$$\mathrm{iSINR}(f) = \frac{M\,\phi_{X_1}(f)}{\mathrm{tr}\,[\mathbf{\Phi}_{\mathrm{in}}(f) + \mathbf{\Phi}_{\mathbf{v}}(f)]}, \tag{G.5}$$

where $\mathrm{tr}\,[\cdot]$ denotes the trace of a square matrix.

We apply a complex-valued filter, or a beamformer as we refer to, $\mathbf{h}(f) = [\,H_1(f)\ H_2(f)\ \cdots\ H_M(f)\,]^{\mathrm{T}}$ on the microphone outputs, that results $Z(f) = \mathbf{h}^{\mathrm{H}}(f)\,\mathbf{y}(f)$ with the variance of

$$\begin{aligned}\phi_Z(f) = \mathbf{h}^{\mathrm{H}}(f)\,\mathbf{d}_1(f)\,\phi_{X_1}(f)\,\mathbf{d}_1^{\mathrm{H}}(f)\,\mathbf{h}(f) + \\ \mathbf{h}^{\mathrm{H}}(f)\,[\,\mathbf{\Phi}_{\mathrm{in}}(f) + \mathbf{\Phi}_{\mathbf{v}}(f)\,]\,\mathbf{h}(f).\end{aligned} \tag{G.6}$$

With the distortionless constraint that $\mathbf{h}^{\mathrm{H}}(f)\mathbf{d}_1(f) = 1$, we can write the narrowband output SIR and output SINR respectively like

$$\mathrm{oSIR}[\mathbf{h}(f)] = \frac{\phi_{X_1}(f)}{\mathbf{h}^{\mathrm{H}}(f)\,\mathbf{\Phi}_{\mathrm{in}}(f)\,\mathbf{h}(f)}, \tag{G.7}$$

$$\mathrm{oSINR}[\mathbf{h}(f)] = \frac{\phi_{X_1}(f)}{\mathbf{h}^{\mathrm{H}}(f)\,[\,\mathbf{\Phi}_{\mathrm{in}}(f) + \mathbf{\Phi}_{\mathbf{v}}(f)\,]\,\mathbf{h}(f)}. \tag{G.8}$$

2.2 Minimum Variance Beamformers

A fixed beamformer is a signal independent filter with a specific beampattern, e.g., the DS beamforming has a unit gain at the desired DOA, i.e., $\mathbf{h}_{\mathrm{DS}}(f) = \mathbf{d}_1(f)/M$. However the desired signal is obtained from the desired direction, the output signal suffers from interference-plus-noise except for the unlikely cases when the nulls of the DS beamformer are situated at the direction of interferers. Signal dependent beamformers are designed adaptively to minimize the variance of the output signal. The MVDR or the Capon method [11] minimizes the output interference-plus-noise variance of the beamformer [20], i.e.,

$$\begin{aligned} \min_{\mathbf{h}(f)} \quad & \mathbf{h}^{\mathrm{H}}(f)\,[\,\mathbf{\Phi}_{\mathrm{in}}(f) + \mathbf{\Phi}_{\mathbf{v}}(f)\,]\,\mathbf{h}(f) \\ \text{subject to} \quad & \mathbf{h}^{\mathrm{H}}(f)\,\mathbf{d}_1(f) = 1, \end{aligned} \tag{G.9}$$

and the MVDR beamformer is given by [4]

$$\mathbf{h}_{\mathrm{M}}(f) \;=\; \frac{[\,\mathbf{\Phi}_{\mathrm{in}}(f) + \mathbf{\Phi}_{\mathbf{v}}(f)\,]^{-1}\mathbf{d}_1(f)}{\mathbf{d}_1^{\mathrm{H}}(f)\,[\,\mathbf{\Phi}_{\mathrm{in}}(f) + \mathbf{\Phi}_{\mathbf{v}}(f)\,]^{-1}\mathbf{d}_1(f)}. \tag{G.10}$$

In the MVDR filter design, interferers are assumed to be uncorrelated with the desired signal; otherwise the desired signal may be suppressed. Herein, we generalize the MVDR beamformer to derive the LCMV filter that nulls out $N-1$ number of interferers and minimizes the noise variance, i.e.,

$$\begin{aligned} \min_{\mathbf{h}(f)} \quad & \mathbf{h}^{\mathrm{H}}(f)\,\mathbf{\Phi}_{\mathbf{v}}(f)\,\mathbf{h}(f) \\ \text{subject to} \quad & \mathbf{h}^{\mathrm{H}}(f)\,\mathbf{D}(f) = \mathbf{i}_N^{\mathrm{T}}, \end{aligned} \tag{G.11}$$

where $\mathbf{i}_N$ is the first column of the $N \times N$ identity matrix, $\mathbf{I}_N$. The solution for the LCMV beamformer is

$$\mathbf{h}_{\mathrm{L}}(f) = \mathbf{\Phi}_{\mathbf{v}}^{-1}(f)\,\mathbf{D}(f)[\,\mathbf{D}^{\mathrm{H}}(f)\,\mathbf{\Phi}_{\mathbf{v}}^{-1}(f)\,\mathbf{D}(f)\,]^{-1}\mathbf{i}_N. \tag{G.12}$$

3 Proposed Method

The optimization procedures in the MVDR and the LCMV beamformers consist of the number of constraints and the residual (interference-plus-)noise. To design a beamformer which has properties between those beamformers, we now introduce a general expression for the signal model. We divide N signal sources into two sets of N_1 and N_2 sources as $\mathbf{x}(f) = [\,\mathbf{x}_{N_1}^{\mathrm{T}}(f)\;\mathbf{x}_{N_2}^{\mathrm{T}}(f)\,]^{\mathrm{T}}$. Therefore, the received signals can be written like

$$\mathbf{y}(f) = \mathbf{D}_{N1}(f)\,\mathbf{x}_{N_1}(f) + [\,\mathbf{D}_{N_2}(f)\,\mathbf{x}_{N_2}(f) + \mathbf{v}(f)\,], \tag{G.13}$$

where $\mathbf{D}_{N_1}(f)$ and $\mathbf{D}_{N_2}(f)$ are matrices containing the steering vectors of the related signal sets, i.e., $\mathbf{D}(f) = [\,\mathbf{D}_{N_1}(f)\ \mathbf{D}_{N_2}(f)\,]$. We can rewrite the correlation matrix of this decomposition as

$$\mathbf{\Phi}_{\mathbf{y}}(f) = \mathbf{D}_{N_1}(f)\,\mathbf{\Phi}_{\mathbf{x}_{N_1}}(f)\,\mathbf{D}_{N_1}^{\mathrm{H}}(f) + \mathbf{\Phi}_{\mathrm{in},N_2}(f) + \mathbf{\Phi}_{\mathbf{v}}(f), \tag{G.14}$$

where $\mathbf{\Phi}_{\mathrm{in},N_2}(f) = \mathbf{D}_{N_2}(f)\,\mathbf{\Phi}_{\mathbf{x}_{N_2}}(f)\,\mathbf{D}_{N_2}^{\mathrm{H}}(f)$, and $\mathbf{\Phi}_{\mathbf{x}_{N_1}}(f)$ and $\mathbf{\Phi}_{\mathbf{x}_{N_2}}(f)$ are the correlation matrices of the $\mathbf{x}_{N_1}^{\mathrm{T}}(f)$ and $\mathbf{x}_{N_1}^{\mathrm{T}}(f)$ signal sets, respectively.

We apply the signal decomposition model (G.13) to propose a beamformer which we call the controllable LCMV (C-LCMV) inspired from LCMV and MVDR beamformers. For the set of N_1 signal sources, containing the desired signal, the filter is constrained to null out the remaining $N_1 - 1$ interferers, and the remaining $N_2 = N - N_1$ signal sources are minimized together with the background noise, i.e.,

$$\begin{aligned} \min_{\mathbf{h}(f)} \quad & \mathbf{h}^{\mathrm{H}}(f)\,[\,\mathbf{\Phi}_{\mathrm{in},N_2}(f) + \mathbf{\Phi}_{\mathbf{v}}(f)\,]\,\mathbf{h}(f) \\ \text{subject to} \quad & \mathbf{h}^{\mathrm{H}}(f)\,\mathbf{D}_{N_1}(f) = \mathbf{i}_{N_1}^{\mathrm{T}}. \end{aligned} \tag{G.15}$$

The C-LCMV beamformer is designed using the method of Lagrange multipliers as

$$\begin{aligned} \mathbf{h}_{\mathrm{C}}(f) = [\,\mathbf{\Phi}_{\mathrm{in},N_2}(f) + \mathbf{\Phi}_{\mathbf{v}}(f)\,]^{-1}\,\mathbf{D}_{N_1}(f)\times \\ [\,\mathbf{D}_{N_1}^{\mathrm{H}}(f)\,[\,\mathbf{\Phi}_{\mathrm{in},N_2}(f) + \mathbf{\Phi}_{\mathbf{v}}(f)\,]^{-1}\mathbf{D}_{N_1}(f)\,]^{-1}\mathbf{i}_{N_1}. \end{aligned} \tag{G.16}$$

This optimal filter is controlled using a different number of constraints, i.e. $N_1 = 1, 2, \ldots, N$. In particular cases, if $N_1 = 1$ or $N_1 = N$, the filter will be the MVDR beamformer or the LCMV beamformer, respectively. Therefore, the C-LCMV beamformer has the following properties:

$$\mathrm{oSINR}\,[\mathbf{h}_{\mathrm{L}}(f)] \leq \mathrm{oSINR}\,[\mathbf{h}_{\mathrm{C}}(f)] \leq \mathrm{oSINR}\,[\mathbf{h}_{\mathrm{M}}(f)]\,, \tag{G.17}$$

$$\mathrm{oSIR}\,[\mathbf{h}_{\mathrm{M}}(f)] \leq \mathrm{oSIR}\,[\mathbf{h}_{\mathrm{C}}(f)] \leq \mathrm{oSIR}\,[\mathbf{h}_{\mathrm{L}}(f)]\,. \tag{G.18}$$

4 Simulation Results

We investigate the performance of the C-LCMV beamformer comparing with the DS, MVDR, and LCMV beamformers in an anechoic environment. We use a uniform linear array (ULA) which the distance between microphones is δ = 0.04 m, i.e., smaller than the half of the minimum wavelength to avoid spatial aliasing, and the wave propagation speed is assumed c = 340 m/s. By selecting the first microphone as the reference microphone, the steering

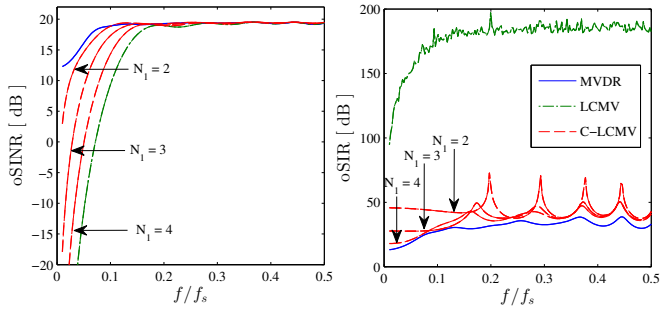


Fig. G.1: Output SINR (left) and output SIR (right) of different beamformers versus frequency (input SINR = 8 dB, and input SIR = 13 dB).

vector $\mathbf{d}_n(f) = \mathbf{d}(f, \theta_n)$ can be written as a function of the DOA of the nth signal source, i.e., θ_n, as

$$\mathbf{d}_n(f) = \left[1 \;\; e^{-j2\pi f \tau_0 \cos\theta_n} \cdots \; e^{-j2(M-1)\pi f \tau_0 \cos\theta_n}\right]^{\mathrm{T}}, \tag{G.19}$$

where $j = \sqrt{-1}$, and $\tau_0 = \delta/c$ is the delay between two successive sensors at the zero angle.

In Figure G.1, we plot narrowband oSINRs and oSIRs for various number of constraints N_1, where $M = 9$, and $N = 5$ white Gaussian signal sources at $\theta_1 = \pi/6$, $\theta_2 = \pi/2$, $\theta_3 = 2\pi/3$, $\theta_4 = 5\pi/6$, and $\theta_5 = \pi$. This figure illustrates that the C-LCMV beamformer performs in the range between the MVDR and LCMV beamformers. In the next experiments, we use three speech signals and white Gaussian noise, which are located at θ_n (for $n = 1, 2, 3$, and 4), and synthesized according to the signal model (G.1). The desired speech signal is an utterance of "Then, the sun shine", and interferers are utterances of "Why were you away?" and "Somebody decides to break it!".

The speech signals were sampled at $f_s = 8.0$ kHz during 1.28 sec. The desired speech signal is expected to be enhanced using the aforementioned filters in frequencies 0.1– 4.0 kHz, because the linear constrained beamformers may have a low output SNR at low frequencies [18]. We divide this multi-channel signal into 75% overlap frames with 256 samples, and transform them into frequency-domain using a 256-point discrete Fourier transform (DFT). Finally, the output signal of designed filters are transfered into time-domain using the inverse DFT.

The minimum output power beamformer is closely related to the minimum variance beamformer with the distortionless constraint and the perfect signal match [21]. Therefore, the (interference-plus-)noise correlation matrix can be replaced by $\mathbf{\Phi}_{\mathbf{y}}(f)$ in the filter designs (G.10), (G.12), and (G.16). We run simulations using different number of microphones and background

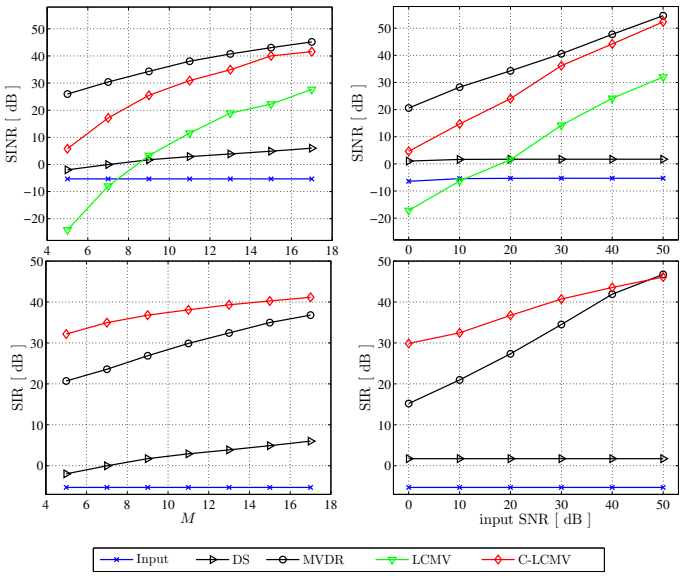


Fig. G.2: Output SINR (top row) and output SIR (bottom row) of different beamformers versus number of microphones in 20 dB noise (left column) and versus input SNR level using $M = 9$ (right column).

noise levels. Since two interfering speech signals may have correlation with the desired signal, that is likely, in the C-LCMV beamformer we only null them out and minimize the power of the uncorrelated interfering signal by choosing $N_1 = 3$. Figure G.2 shows that the broadband oSINR and oSIR of the C-LCMV beamformer performs in the range between the MVDR and LCMV beamformers.

The expectation is estimated by time averaging, and the correlation matrix of the received signals, at a time instance t, is estimated as

$$\hat{\mathbf{\Phi}}_{\mathbf{y},t}(f) = \frac{1}{B}\sum_{b=1}^{B} \mathbf{y}_{t,b}(f)\,\mathbf{y}_{t,b}^{\mathrm{H}}(f), \tag{G.20}$$

where $\mathbf{y}_{t,b}(f)$ is the bth spectral amplitude estimate out of the last B estimates [22]. Moreover, the full rank correlation matrix can be guaranteed by choosing the buffer size as $B \geq M$, and we choose $B = 100$. In practice, the correlation matrix estimate may has error due to the limited number of samples in low iSNRs and the dominant desired signal. Diagonal loading [14] is a solution for this problem, i.e., $\hat{\mathbf{\Phi}}_{\mathbf{y}}(f) \leftarrow \hat{\mathbf{\Phi}}_{\mathbf{y}}(f) + \gamma \mathbf{I}_M$, that we choose $\gamma = 10^{-4}$. In -5 db broadband iSINR (20 dB background noise), Figure G.3 shows spectrograms of the noisy signal at the first microphone, the output signals of beamformers using $M = 11$ microphones. Although the LCMV beamformer outperforms the MVDR beamformer by removing interferers,

the LCMV beamformer distort the speech signal at low frequencies. The experiment results indicate that the C-LCMV beamformer removes interference tracks from the noisy signal without distorting the desired signal at low frequencies.

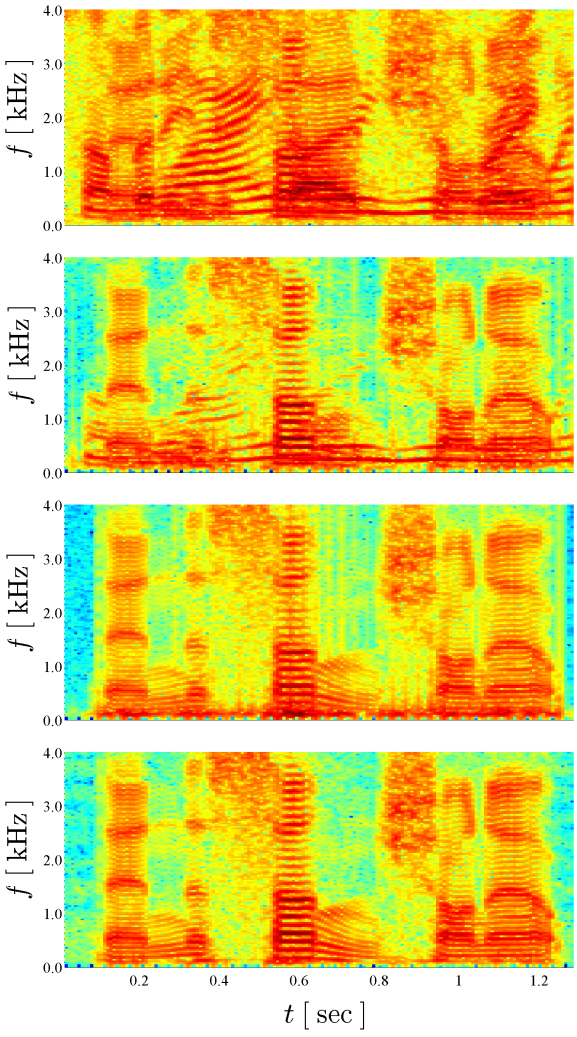


Fig. G.3: According to the order of plots from top to down: the spectrograms of the noisy signal at the first microphone, the output signals of the MVDR, LCMV, and C-LCMV beamformers.

5 Conclusion

The work presented in this paper has focused on signal enhancement in the presence of interference. The LCMV beamformer may have infinite output SIR, but have a lower output SNR than the MVDR beamformer. This problem is increased dramatically using a high number of constraints to remove interferers, especially at low frequencies and closely spaced interference [18]. We have proposed the C-LCMV beamformer being able to control the quality of the signal of interest, a trade-off between noise reduction and interference rejection.

References

[1] J. Benesty, J. Chen, Y. Huang, and I. Cohen, *Noise Reduction in Speech Processing*. Springer-Verlag, 2009.

[2] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction Wiener filter," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, no. 4, pp. 1218–1234, Jul. 2006.

[3] B. D. Van Veen and K. M. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE ASSP Mag.*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[4] J. Benesty, Y. Huang, and J. Chen, *Microphone Array Signal Processing*. Springer-Verlag, 2008, vol. 1.

[5] Y. Kaneda and J. Ohga, "Adaptive microphone-array system for noise reduction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 6, pp. 1391–1400, 1986.

[6] S. Doclo and M. Moonen, "On the output SNR of the speech-distortion weighted multichannel Wiener filter," *Signal Processing Letters, IEEE*, vol. 12, no. 12, pp. 809–811, 2005.

[7] S. Applebaum and D. Chapman, "Adaptive arrays with main beam constraints," *IEEE Trans. Antennas Propag.*, vol. 24, no. 5, pp. 650–662, 1976.

[8] K. Buckley, "Broad-band beamforming and the generalized sidelobe canceller," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 5, pp. 1322–1323, Oct. 1986.

[9] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Signal Process.*, vol. 49, no. 8, pp. 1614–1626, 2001.

[10] H. Cox, R. Zeskind, and T. Kooij, "Practical supergain," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 3, pp. 393–398, 1986.

[11] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.

[12] O. L. Frost, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, vol. 60, no. 8, pp. 926–935, Aug. 1972.

[13] M. Brandstein and D. Ward, Eds., *Microphone Arrays - Signal Processing Techniques and Applications*. Springer-Verlag, 2001.

[14] H. Cox, "Resolving power and sensitivity to mismatch of optimum array processors," *J. Acoust. Soc. Am.*, vol. 54, no. 3, pp. 771–785, Sep. 1973.

[15] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[16] J. Benesty, J. Chen, and E. A. P. Habets, *Speech Enhancement in the STFT Domain*. Springer, 2012, vol. 5.

[17] H. Steyskal, "Wide-band nulling performance versus number of pattern constraints for an array antenna," *IEEE Trans. Antennas Propag.*, vol. 31, no. 1, pp. 159–163, Jan 1983.

[18] M. Souden, J. Benesty, and S. Affes, "A study of the LCMV and MVDR noise reduction filters," *IEEE Trans. Signal Process.*, vol. 58, no. 9, pp. 4925–4935, Sept. 2010.

[19] K. Bell, Y. Ephraim, and H. Van Trees, "A bayesian approach to robust adaptive beamforming," *IEEE Trans. Signal Process.*, vol. 48, no. 2, pp. 386–398, Feb 2000.

[20] H. Cox, R. Zeskind, and M. Owen, "Robust adaptive beamforming," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 35, no. 10, pp. 1365–1376, Oct. 1987.

[21] H. L. Van Trees, *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*. John Wiley & Sons, Inc., 2002.

[22] M. E. Lockwood and et al., "Performance of time- and frequency-domain binaural beamformers based on recorded signals from real rooms," *J. Acoust. Soc. Am.*, vol. 115, no. 1, pp. 379–391, Jan. 2004.

ISSN (online): 2246-1248
ISBN (online): 978-87-7112-750-8

AALBORG UNIVERSITY PRESS